DOCUMENT RESUME

ED 180 770 SE 029 398

AUTHOR Rabins, Michael J.
TITLE An Introduction to Dynamic Systems and Feedback.
INSTITUTION New York Polytechnic Inst., N.Y.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Jul 73
GRANT NSF-GY-2467
NOTE 76p.; Not available in hard copy due to small print throughout entire document

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Engineering Education; Engineers; *Higher Education; Mathematics Education; *Models; Physics; *Science Education; Social Sciences; *Systems Analysis; *Systems Concepts

ABSTRACT

This introduction to dynamic systems is intended for freshman and sophomore students in engineering, physical science, or social science. Material has been class tested and has led to increased student interest in further work in systems analysis and operations research. Notes are written for the student and are self-contained. Material can be covered in about 15 lectures and should be followed by additional problem work. Topics covered include input/output causality, causality and problem formation modeling, solution of first order equations, second order systems, and closed loop feedback systems. Appendices provide additional information. (Author/RE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

REC'D MIDS/HES AUG 20 1973

GY-2467
(EN)

MIDS/HES
ARCHIVES

ED180770

AN INTRODUCTION TO DYNAMIC SYSTEMS AND FEEDBACK

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Mary L. Charles
NSF
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Michael J. Rabins
Department of Operations Research and Systems Analysis
Polytechnic Institute of New York

Sponsored by the National Science Foundation under Grant No. GY-2467
to Prof. Fred Landis of New York University

2

July 1973

SE 029 398

8/20/73 — 10 cys ached by card. GJ

AN INTRODUCTION TO

DYNAMIC SYSTEMS AND FEEDBACK

Michael J. Rabins
Department of Operations Research and System Analysis
Polytechnic Institute of New York

This material was prepared under Grant No. GY–2467 from the National Science Foundation

July, 1973

This booklet is one of a series to provide resource material for Freshman and Sophomore courses in Engineering coordinated by Professor Fred Landis, Department of Mechanical Engineering, New York University*, who serves as Project Director.

(*) the Polytechnic Institute of New York as of September 1973

© New York University 1973

Additional copies may be obtained at nominal cost by contacting Professor Fred Landis, Department of Mechanical Engineering, Polytechnic Institute of New York, 333 Jay Street, Brooklyn, N.Y. 11201

EDITOR'S PREFACE

The study of dynamic response and feedback has become a driving force in modern concepts of systems analysis. The field has moved from its engineering origins and has begun to make its impact felt in the social, behavioral and biological sciences. System dynamic response and feedback are rapidly becoming one of the key bridges in dealing with the interaction between technology and society.

The reasons for this development are not hard to find. Dynamic response takes cognizance of the fact that we live in a rapidly and continuously changing world with many interrelations, most of which are coupled by feedback mechanisms. It shows that relatively simple mathematical models can provide at least initial insight in a large number of problems. It exudes excitement that readily captures the imagination of many college students. And, finally, systems analysis, if properly presented, can develop basic concepts without a premature screen of advanced mathematics.

This introduction to dynamic systems is primarily intended for freshman and sophomore students in engineering and the physical and social sciences. The material was prepared as part of a National Science Foundation sponsored project under the direction of the editor to develop resource material for lower level engineering students. The material in these notes has been class-tested in a freshman course at the Polytechnic Institute of Brooklyn where it has also led to increased student interest in further work in systems analysis and operations research.

The notes are written for the student and are self contained. However, further examples brought in by an instructor would be highly desirable. The material can be covered in about fifteen lectures and should be followed by additional problem work dealing with technology, ecology and urban planning. This has been the pattern at the Polytechnic Institute of Brooklyn where the notes form the early part of a course "Concepts of Modern Technology" which is described more fully in the appendices. Alternatively, the material can be supplemented by work in operations research. Here the booklet "An Introduction to Operations Research" by Professor L. Graff, which was released earlier as part of this NSF sponsored program, may provide the instructional base.

As in many recent books on technological concepts, these notes owe a debt of gratitude to the series "The Man-Made World" and "Man and His Technology" which were compiled under the auspices of the Engineering Concepts Curriculum Committee. The instructor may well wish to have recourse to these volumes in fleshing out his presentation. Many laboratory experiments directly applicable to these notes can be drawn directly from "The Man-Made World". However, the primary focus of these notes is different. Instead of providing a broad range first coverage, the material focuses on one primary topic by presenting the basic ideas of mathematical modeling and analysis of dynamic systems, the concepts of errors and gains, and, finally, the qualitative effects of feedback on various engineering and social systems. This can be done on the freshman level and the feeling of understanding and achievement gained by the student becomes a strong motivation for further engineering studies.

Fred Landis, Chairman
Mechanical Engineering Department
New York University
Project Director

July 1973

Instructor's Preface

This set of notes grew out of lecture material developed for a new course on Concepts of Modern Technology which is directed towards students in engineering, the physical and the social sciences. The course aims to present an introductory engineering point of view and to relate technology to the problems of society.

Any course based on these notes should keep two main objectives in mind. First, students embarking on a career in science or technology should become aware of the role that technology plays in their daily lives, and develop a sensitivity to societal problems. Secondly, it should provide the non-technical student with a glimpse of technological perspectives and show how an engineering approach becomes a valuable tool in coping with the problems of society.

These notes are based on actual class lectures at the Polytechnic Institute of Brooklyn in New York City. Accordingly, the material is student oriented and this preface should serve as an explanation of a procedure that has worked in practice. At Brooklyn Poly the Concepts of Modern Technology course is required of all freshmen majoring in the System Engineering option of the Department of Operations Research and System Analysis. It may also be taken as an elective by other engineering, mathematics, social science and physical science majors.

The rationale in offering this course is to introduce the student to some of the many important concepts and techniques to be learned in the ensuing years of college and to present an overview and a motivational introduction to what comes later in the curriculum. The course is a self-standing entity and requires no mathematics background on the college level although familiarity with FORTRAN programming is assumed.

The class meets for one weekly two-hour lecture period over the 14 week term. Appendix A shows the course syllabus used during 1972. These notes cover the first seven weeks of that syllabus plus selected portions of subsequent material. In addition, the class is scheduled to meet 3 hours per week in a workshop-laboratory. In these workshops, for the first half of the semester, the class investigates analog and digital simulations. Information regarding the digital and analog computer workshops is listed in Appendix B. The digital simulation programs listed in this appendix are in BASIC language and are performed on PDP–8 time sharing terminals. They were developed at the Polytechnic Institute under the NSF sponsored "Huntington Project." Details about this project and the programs are available from Prof. L. Braun of the State University of New York at Stony Brook, Long Island, N. Y. Descriptions of the analog simulations may be found in the "Laboratory Assignments" at the back of chapters four through nine in the course text, The Man-Made World (McGraw-Hill, 1971). For these analog workshops, reference to the text is advisable, but the lecture material contained in these notes is self-sufficient. Although the workshop and lectures are essentially independent, they are strongly coupled and coordinated in the Brooklyn Poly format actually used.

During the second half of the semester the students work on 2-man team projects of their own choosing. Typical past projects include analog simulation of a parachute jump, digital simulation of a baseball game, and analog simulation of black population shift in a neighborhood. These projects, with both faculty and graduate student advisers guiding them, culminate in written reports at the end of the semester. Use of such individually selected projects has proven an effective part of the course.

Students are also supplied a reading list (see Appendix C) at the outset of the course. Study of this list indicates that the focus is upon social impact of technological change and the consience of technologists. Students are instructed to choose two books from this list and to report on them in written form and in a class seminar presentation later in the semester. While the reading list is not directly related to the present notes, the list provides a frame of reference and perspective which is easily introduced into the lectures throughout the course.

Although the material presented in the following chapters is based mainly upon the lecture portion of the Polytechnic course "Concepts of Modern Technology," it also draws heavily upon experience gained in a summer institute sponsored by the National Science Foundation which was held during the past three summers at the Polytechnic. This institute, entitled "Technology and Society – System Analysis and Simulation" and offered under the aegis of the Science Student Training Program (SSTP), is directed at high-ability students between their third and fourth year of high school. Details about this NSF summer program are available from the author.

The author is indebted to Professor Fred Landis of New York University for his many helpful discussions during the early versions of the manuscript and for his corrections in the final version. Whatever coherence and perspective have emerged should be largely attributed to him.

The typing of these notes was done by Ms. Marsha Rabinowitz and Mr. Fred Burg. The art work is due to Mr. Flavio Dobran. Their help and diligence are gratefully acknowledged.

Michael J. Rabins
Professor and Program Director of System Engineering
Department of Operations Research and System Analysis
Polytechnic Institute of New York
July, 1973

Table Of Contents

	Page
Editor's Preface	i
Instructor's Preface	ii
Chapter 1. Input/Output Causality Introduction – Types of Causality – Accumulation – Computation of Accumulation – Analog Simulation	1
Chapter 2. Causality and Problem Formulation Modeling Introduction – Problem Formulation, An Example – Constraints – Mathematical Models – Criteria, Alternatives and Tradeoffs – Mathematical Models and Causality – Linearization and Modeling	7
Chapter 3. The Solution of First Order Equations Introduction – Numerical Solution – Analog Computer Solution – Mathematical Analysis	17
Chapter 4. Second Order Systems Introduction – Second Order System Models – Analog Simulation of Second Order Systems – Analysis of Second Order Systems – Numerical Analysis of Second Order Systems – Higher Order Systems	31
Chapter 5. Closed Loop Feedback Systems Introduction – Definitions – Historical Background – Introduction to Feedback Control System Analysis – An Example of a Typical Feedback Control System – Closure	51
Appendix A. Course Syllabus of "Concepts of Modern Technology"	65
Appendix B. Freshman Workshop Assignments; Digital and Analog Computer Simulation	66
Appendix C. Reading List for "Concepts of Modern Technology" Course	68

CHAPTER 1

INPUT/OUTPUT CAUSALITY

Introduction

It is clear that we live in a world which becomes more and more technologically oriented. In our daily existence we can observe the impact on our style of living made by advances in communications, transportation, health care and computer science, to name just a few areas. An engineer or scientist working in these fields must be sensitive to these impacts on the world around him. At the same time, others not working in technology or science should gain an appreciation of the detailed role that technology plays in changing our social fabric. The goal of this set of notes is to exemplify one approach towards developing such a two-way sensitivity and appreciation.

As a first step, we shall examine the interrelation between cause and effect; which is a cause, which is an effect, what leads from one to the other and how. While these first steps may initially appear small or obvious, the perspective that is introduced will be necessary to handle more complex problems later on.

Types of Causality

We are all familiar with cause and effect relations. In our daily lives we are constantly in situations where an input, or cause, is required to yield an output, or effect. When we open the faucet in a sink with the drain plugged, the sink starts to fill up. When we get into an elevator and press a floor button, the elevator moves to the floor we desire. When we push down on one end of a see-saw, a child on the other end rises into the air. When some haphazard and apparently unrelated sequence of events takes place, stock market prices may go up. When the temperature in a room goes up, the liquid in the thermometer in the room rises.

These examples just cited are different types of cause and effect relations and may be broken down into different categories: first, cause and effect relations (input and output) may be either unilateral or reversible. Thus, a rise in room temperature causes a raising of the liquid in the thermometer; not vice-versa. Hence, we have here a unilateral causality where the input always causes the output. An example of reversible causality is seen in the case of the see-saw; our input of a force (cause) which results in the output (effect) of a child rising on the other end of a see-saw, may then be reversed. Now that the child is perched up on the end of the see-saw, he then acts as an input (cause) yielding the output (effect) of a force on us. We see that input and output terminology has no permanent meaning in reversible causality situations.

In addition to unilateral vs. reversible causality, we must also consider deterministic vs. non-deterministic causality. A deterministic system is one in which inputs are completely specified for all time so that the system output may then be completely specified by a knowledgeable analyst. A clear counter-example is the stock market situation cited. The sink faucet case is a possible example of a deterministic case. Once the faucet is opened to a preset amount, with the drain closed, the sink will fill at a predetermined rate assuming a constant city water supply pressure.

There is still another category into which cause and effect relationships may fall: discrete vs. continuous-in-time relationships. In the discrete case, either an occurence takes place or it doesn't. This may be described digitally as an "either-or" statement or by true or false, zero or one, or plus or minus. The elevator button example is a case in point. The continuous-in-time cause and effect relationship, sometimes referred to as the analog siutation, appears when the input exists at all instants of time in a continuous and smooth fashion. In such situations, we can usually expect the system output, or effect, to occur in a smooth and continuous fashion. Examples

of this situation are the water continuously flowing from the faucet and continuously filling the sink, or the continuously rising temperature in the room causing a continuously rising temperature on the thermometer.

In these notes, the latter category of continuous-in-time and deterministic cause and effect relations will occupy the major portion of our time. This is not to belittle the other categories of input/output relationships. However, a study of continuous time-dependent events will be rich in problems taken from our society, and will keep us well occupied.

Accumulation

In order to better understand the input/output behavior of an engineering or societal system, we must at this point introduce the concept of the rate of change of a variable. There are many situation around us in which the rate of change of a variable determines the level of that variable. Thus, the volumetric rate of flow of water from the faucet (in cubic inches per second) continuously determines the level of the volume of water (in cubic inches) accumulated in the stopped sink. The level of water in the sink at any instant (the effect) is directly determined by the rate of flow of water out of the faucet (the cause). Since the rate of flow of water from the faucet is in units of volume per unit time, we therefore denote it as the rate of change of the variable (volume) in question and note that the first directly causes the second. Thus, if the faucet is opened to provide double the flow rate, then the level will rise twice as fast: e.g., if the faucet flow rate is doubled from 1 to 2 cubic inches per second, then, assuming the filling process starts at t=0 with an empty sink, after ten seconds we will have accumulated 10 cubic inches in the first case and 20 cubic inches in the second case.

We need only look around us to observe further examples of the rate of change of a variable causing the change in the level of that variable as an output (effect). If we drive an automobile at constant speed (miles/hour) – which is the rate of change of the distance variable per unit time – then the distance travelled accumulates at a constant rate. Other variables whose rate of change we might examine include: rate of change of population of some locale measured in thousands of people per year, or rate of flow of tons of rough sewage into a river per day, or rate of change of velocity in miles per hour per hour (acceleration), or rate of change of warehouse inventory in units per month, or rate of entry of vehicles into a freeway per hour, etc. In each case, the rate of change of the variable (people, tons, mph, units, vehicles, etc.) leads to the accumulation of the level of that variable within the system under study.

In Figure 1, we draw a plot of some generalized variable "X", vs. time. Using the same time axis we also plot the rate of change of that variable X per unit time. We will use shorthand notation for the rate of change of the variable X and denote it by $\dot{X}$. We observe in Figure 1, that with a constant rate of change of a variable ($\dot{X}$ = constant), the variable X plotted increases linearly with time. Going back to the several examples cited above we can see the physical reasonableness of this concept in that the level, X, accumulates continuously as the rate of change of that level, $\dot{X}$, continues on unabated.

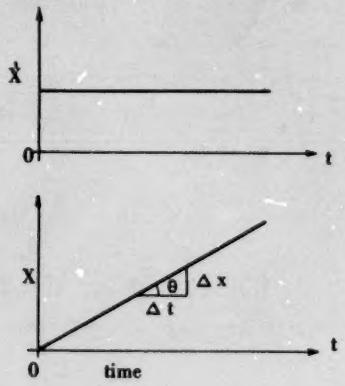


Figure 1

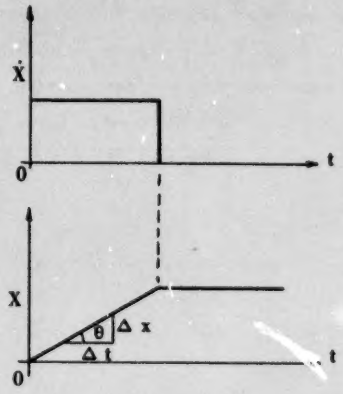


Figure 2

In Figure 2 we observe what happens when the rate of change of a variable, after remaining constant for a while, falls off sharply to zero. We observe in Figure 2 that the level of the variable X rises linearly with time as long as the rate of change of the variable, $\dot{X}$, remains constant. As soon as the rate of change of the variable, $\dot{X}$, goes to zero, no further accumulation of the level of the variable X

occurs and it remains constant. Examples here include turning off the faucet so that the level in the sink remains constant, or 'preventing' all emigration, immigration, births and deaths in a locale under study so that the population remains constant.

A further complication exists since the rate of change of a variable may be negative. Thus, for example, we may open the stopper of the sink and let water flow out. Or we may turn a car around and drive in the opposite direction (with negative velocity), thus, accumulating negative distance. In this case, we are just heading back to where we started. Figure 3 shows a generalized example of a rate variable, $\dot{X}$, which starts out positive, becomes zero for a while, and then ends up negative. The resulting plot of the variable X vs. time may be easily traced based upon the previous discussion. In Fig. 3, note that the negative value of $\dot{X}$ is (absolutely) larger than the initial positive value, so that the variable X accumulates slower at the start than it disperses at the end.

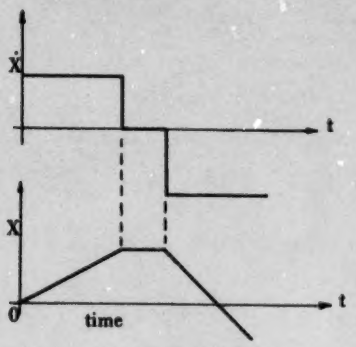


Figure 3

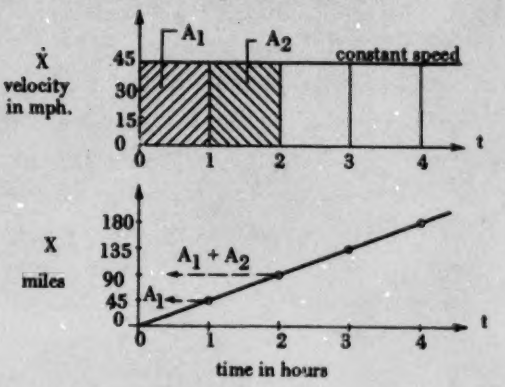


Figure 4

It is interesting to note that the area under the rate vs. time plot, up to any instant of time, is precisely equal to the level of the variable accumulated to that instant. Thus, in Figure 4, with a plot of velocity in miles per hour vs. hours of elapsed time, we note that after two hours at a constant velocity of 45 mph the automobile has travelled 90 (45 mph times 2 hours) miles. Note carefully that the units of hours in numerator and denominator cancel out to yield the correct dimension of miles in the accumulated distance of 90 miles after two hours. It is an easy matter to check other points of the level curve of distance travelled in the same manner. Thus, while the distance travelled in two hours is given by the area $A_1 + A_2$ in Fig. 4, the differential distance travelled in the second hour is just equal to the area A_2.

Still another complication occurs if the rate of change of a variable does not remain constant. It may be increasing or decreasing in time, or it may be oscillating in time or it may be behaving in a random fashion. In Figure 5 we observe what occurs when the rate of change of a variable, $\dot{X}$, is increasing linearly with time. We assume that at t=0 our variable, X, started with an initial value of zero. We note that the level variable X in this case, increases at an ever increasing rate. The points on the level curve were obtained by computing the area underneath the rate curve after the selected intervals of time.

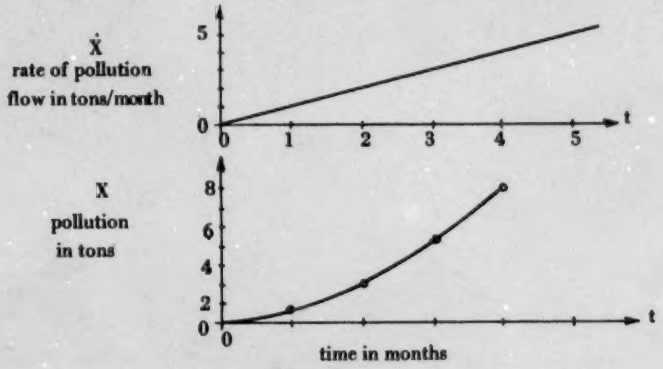


Figure 5

Thus, in Figure 5, if our variable is the amount of pollution (tons of sewage perhaps) in some body of water and the rate of change of the pollution is increasing linearly with time as shown in the top portion of the figure, then we may easily compute the amount of pollution after any interval of time. Referring to Figure 5, we see that after one month the rate of change of pollution has reached a value of one ton per month. The area underneath the rate of change plot from t = 0 to t = 1 month is the area underneath the triangle defined by these points and equals one half of a ton of pollution. This is the amount of accumulated pollution plotted in the level curve of pollution immediately underneath the rate of change of pollution curve. Since the rate of pollution flow is increasingly linearly, after two months we are therefore creating pollution at the rate of two tons per month. The area under the larger traingle between zero and two months represents two tons of accumulated pollution. After three months, it is seen that 4½ tons of pollution have accumulated. After 4 months it is seen that 8 tons of pollution have accumulated. The result is a quadratically growing accumulation of pollution.*

It is not difficult to conceive of a case where the rate of change of a variable in changing markedly. And yet, by applying the concept discussed above of computing accumulated level by measuring area under the rate curve, we are able, even for complex cases, to arrive at a level vs. time plot. Thus, in Figure 6, the vertical acceleration (rate of change of velocity) of a rocket leads to the plot of the velocity vs. time by computing areas underneath the acceleration plot. Note here that we are dealing with the time rate of change of velocity – which must be considered at this point as a level of accumulated velocity. Similarly, computing areas underneath the velocity (rate of change of distance) plot vs. time leads to the accumulation of distance travelled by the rocket. We now begin to understand the importance of accurate accelerometers which can measure the acceleration of a rocket and transmit that information back to earth. Note that after the rocket burns our (thrust goes to zero and therefore so does acceleration by Newton's second law), and assuming the rocket has escaped any gravitational pull, then the rocket velocity remains constant and distance accumulates continuously and linearly with time as the rocket moves off in outer space.

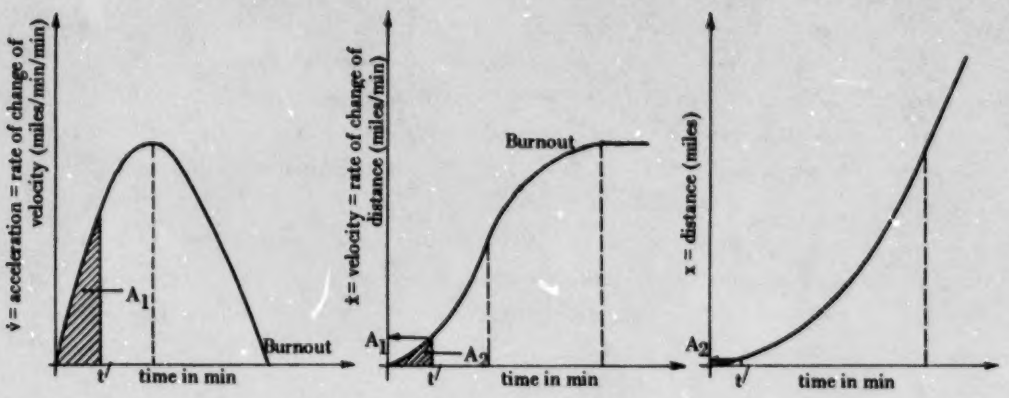


Figure 6. Acceleration, Velocity and Distance vs. Time for a Rocket.

Computation of Accumulation

Let us now return to the simpler case of Figures 1 and 2 in which the rate of change of the variable remains constant, or in other words, $\dot{X}$ = constant. Let us introduce, at this point, a new notation utilizing the Greek letter delta, Δ, which will signify "change in." The rate of change of the variable X per unit time will therefore be written as,

$$\frac{\Delta X}{\Delta t} \equiv \dot{X} . \qquad (1)$$

* This result is not surprising to the student of integral calculus. The heading of this section, "accumulation", is here used synonymously with the term integration. This leads to the phrase "integral causality": i.e., in all of the examples examined thus far, it is the rate of change of something that must be accumulated, or integrated, to yield a change in that something. In general, we do not find the opposite situation in nature or society wherein a level might act as in input to cause a change in the rate of change of that level. This is denoted "differential causality" and is an unnatural causality which does not occur in real situations of concern to us.

In Figures 1 and 2 we note that the quantity $\Delta X/\Delta t$ is constant and is representative of the slope (= tangent θ) of the variable plot vs. time. Further we note that the slope of the X vs. time plot is constant, which checks with the fact that the plot of the rate of change of the variable, $\dot{X}$, remains constant. Utilizing this concept of change (Δ) in X per unit time we may now develop a computational scheme. If we know the value of X at some initial starting point and we continuously know the change in X per unit time, then we should be able to compute the level of X after any elapsed time. Calling the constant value of $(\Delta X/\Delta t) = C$ and letting the initial value of X be X_0, then the above description of the situation may be summarized as:

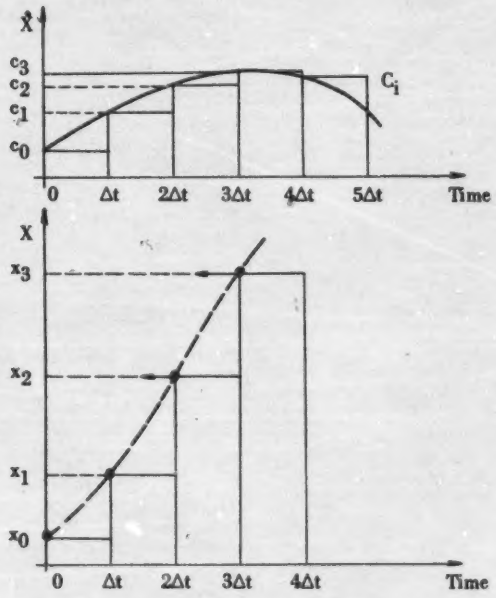
$$\frac{\Delta X}{\Delta t} = C$$

$$X(t) = X(t) - X_0 = C\Delta t$$

or $$X(t) = X_0 + C\Delta t. \qquad (2)$$

We note in Eq. 2 that the quantity X(t) signifies the level of the variable X as a function of time (t). We further note that knowing C, X_0 and Δt, we are able to compute the variable X after any elapsed Δt. Note carefully that the initial value of X is needed in Eq.(2) to perform any calculations.

Now this simple computation scheme is satisfactory as long as the rate of change of the variable remains constant, or nearly so. However, when the rate of change of the variable under study is fluctuating largely with time, then a new approach must be utilized. In this case $\dot{X}$ is only approximately equal to $\Delta X/\Delta t$. We now can introduce the concept of a staircase approximation to a curve as shown in Figure 7., to handle a time varying input rate signal. It should be obvious that the computational scheme then requires repeated application of Eq.(2) as one "climbs the stairs", using Eq.(2) to compute the value of X at the end of each interval Δt.



Note that X increases as long as $\dot{X}$ is positive. For increasing $\dot{X}$ slope of X increases while for decreasing $\dot{X}$ (after $3\Delta t$) slope of X decreases.

Figure 7

The accuracy of this approach will depend, of course, on how well we approximate the time varying signal with our staircase. A better approximation would require a smaller Δt computation interval. (In the limit, as $t \to 0$, $\Delta X/\Delta t$ will exactly equal $\dot{X}$ at all instants as long as the rate curve is continuous.) The application of this approach to compute the level of the variable X_i at T_i in Figure 7 would therefore appear as,

$$X_i = C_{i-1}(\Delta t_i) + X_{i-1}, \qquad (3)$$

where C_{i-1} is the value of X in the interval just before the instant in question. This computational algorithm is easily programmed on a digital computer and is generally known as the Fuler integration method. In this recursive computational scheme, the value computed at each interval becomes the initial value (X_0 in Equation 2) for the next interval. The smaller Δt chosen for non-constant rates, then the more accurate will be the integration results.

Analog Simulation

It is also possible to perform an analog simulation of the solution of the equation (2). We need only conceive of an electronic device, that when fed a constant voltage, puts out a linearly increasing voltage. Such a device is called the "integrating operational amplifier" of an analog computer.* It is not important at this point to understand precisely how such a device operates. Suffice it to say, that the output voltage of the integrator in an analog computer will behave just like the accumulated level figures plotted earlier, when the voltage input to the unit behaves like the rate curves given, (i.e., the unit accumulates the area under the rate curve). With this in mind the stage is set for further simulation work in Chapter 3.

* At this point the large scale demonstration unit of an integrating operational amplifier is shown to the class. Such a unit has been developed by the Engineering Concepts Curriculum Program and is available at reasonable cost from the American Machine and Foundry Co., Elec. Prod. Development Division, 1025 N. Royal St., Alexandria, Va. 22314. However, any operational amplifier with a suitable output device may be used.

CHAPTER 2

CAUSALITY AND PROBLEM FORMULATION MODELING

Introduction

When assigned a new job, the first task should be to carefully define or formulate the problem. In technological problems this first step more often than not leads to a mathematical model of the problem situation. Such a model can take various forms, some of which will be discussed in this chapter, but in every case the detailed modeling steps are a direct outgrowth of the specific goal or objective selected.

Problem Formulation -- An Example

To underline this last thought, let us consider various different and possible problem formulations for the general traffic situation on the Island of Manhattan in New York City. If our goal is the prediction of the number of cars in Manhattan (a level) as a function of time during the day, then a single level variable corresponding to that number would be sufficient for the model at any instant of time. Feeding this level would be rates of flow of automobiles into and out of Manhattan. These rates might be constants, functions of time, or even functions of the level itself. Alternatively, our assignment might be the prediction of potential traffic jam locations and times. Clearly, a single variable model would no longer be sufficient to describe the problem and the number of cars at various locations and during various times of the day would be required for this more detailed study. This now becomes a multi-variable mathematical model with a different variable assigned to the number or level of cars at each specific location and at any time. If the problem be further complicated by predicting the effect of banning private automobiles from certain streets or avenues, a proper traffic study would have to build on the previous model but add the variables associated with the use of subways, buses and taxis which would be additional level variables. It should be obvious that different goals will lead to entirely different problem formulations and mathematical models.

Constraints

Often, in defining the variables of a mathematical model, we must include in our problem definition constraints on both the rates and the levels of the model. Thus, in the first example cited above predicting the total number of cars in Manhattan as a function of time during the day, we might conceive of the following two constraints on our problem formulation:

a) The rate of flow of automobiles into or out of Manhattan at any instant of time is constrained (limited) by the capacity of the major bridges, tunnels and thoroughfares to handle this rate.

and b) The total level of automobiles in Manhattan at any instant of time must not exceed the capacity of all of the parking facilities and streets to accomodate that level assuming cars are spaced normally on the streets and are parked bumper to bumper on every available square foot. We might even wish to tighten this constraint by immediately assuming that a certain fraction of the available parking spaces will not be used in practice and that a tight parking situation in the southern Wall Street area may not be resolved by excess parking space at the northern end of the island about 15 miles away.

These constraints are not merely limiting the bounds of applicability of the solution we are seeking . . . they also often dictate the form of the problem formulation and the steps to be taken in arriving at a solution. Let us consider, for example, the two alternative

models of traffic accumulation in Manhattan for the saturated and unsaturated rate of flow cases.

Mathematical Models

First taking the saturated case where the rate of flow of autos is at a maximum constant value C, for the level of automobile X(t), at any instant of time (t) we get :

$$\Delta X(t)/\Delta t = C, \qquad (4)$$

where $\Delta X(t)$* is the change in the level of automobiles per Δt time change. This ratio is the flow rate which is constrained to be at the maximum (saturated) value of C cars per hour. Equation (4) was seen to have the solution given by Equation (2) which was plotted in Figure 1 in the previous chapter. This is simply the case of a constant rate resulting in the linearly increasing accumulation of the level.

Now consider the case of the initial exodus of cars out of Manhattan at the end of a working day. Let us assume here that saturation does not occur and that the bridges and tunnels can handle the traffic without delay. The rate of flow of automobiles out of Manhattan at any instant of time will probably be a function of the number of automobiles still in Manhattan at that time. Mathematically, we write this statement as

$$\Delta X(t)/\Delta t = f[X(t)], \qquad (5)$$

where the notation $f[X(t)]$ signifies some as yet unspecified function of the variable X(t). As a specific example, consider the situation when the rate of flow of automobiles (i.e., the function of Equation (5)), is some constant k times the instantaneous level of automobiles. Equation (5) becomes

$$\Delta X(t)/\Delta t = -k \cdot X(t) \qquad (6)^{**}$$

Equation (6) is quite different in form from Equation (4), and we should expect a markedly different solution. We will show this to be the case in the next chapter when we deal with the solution of equations like Equation (6).

We shall soon see formulations corresponding to Equation (6) cropping up with increasing frequency to describe various and diverse problem situations (e.g., simplified models of population growth, species survival, warehouse inventory, epidemic dynamics, etc.). Mathematically, the general form of Equation (6) is known as a first order linear difference equation with constant coefficients. The phrase first order signifies that the deltas (Δ's) appear only in the first power. Later we will examine higher order equations. If k were itself a function of time, the equation would be denoted as a first order linear difference equation with time-varying coefficients. On the other hand, if k were equal to X(t) at all instants of time, then the right hand side of Equation (6) would be $[X(t)]^2$ and we would have an example of a so-called nonlinear first order difference equation with constant coefficients (unity in this case). An equation is considered linear when the dependent variable X(t) always appears with a unity exponent. In these notes we shall limit ourselves to linear problems with constant coefficients. The reason is simple; there are only a handful of nonlinear problems for which analytical solutions exist, and we would be continually forced to give up on analysis and resort to numerical problem solutions on the digital computer to achieve a result. Fortunately, we will find that the solutions to linear problems agree with physical reality quite often and give us a good insight into the nature of the real world behavior. However, we must note that most real world problems are indeed non-linear so we will have to return to this problem later in this chapter under the heading of linearization.

One additional comment about Equation (6) is necessary regarding the phrase difference equation. Since we are dealing with deltas (Δ's) we are talking about small but finite changes of the variable X(t) in a finite time interval Δt. If, however, the time interval Δt is permitted to become extremely small (the infinitesimally small time interval dt in the limit), then the difference equation would in the limit, become a differential equation written as

$$dX(t)/dt = -k \cdot X(t) \ , \qquad (7)$$

---

* The notation "X(t)" means the dependent variable X as a function of the independent variable t.

** The minus sign on the right hand side of Equation (6) is intentional and serves an important purpose. It implies the assumption of a sign convention for the flow rate of automobiles -- positive for inflow tending to build up the level of X, and negative for outflow tending to deplete the level. Since were are here dealing with a model for the exodus of automobiles we must have the minus sign in Equation (6). Note that this also permits us to work with a positive constant k.

for the case of Equation (6). This result would be termed a first order, linear differential equation with constant coefficients. The calculus affords us some means of solving such equations and in future chapters we shall explore the solution of both the difference and differential equations.

We might generalize Equation (6) slightly to simultaneously handle the inflow of automobiles as well as the exodus. This might be accomplished by adding a term on the right hand side of the equation to get

$$\Delta X(t)/\Delta t = -k \cdot X(t) + F(t) \qquad (8)$$

where F(t) is the positive inflow rate of automobiles in cars per hour and would be some specified function of time. For example, F(t) might be a constant 50,000 cars per hour between the hours of 7:00 and 9:30 A. M. on an average working day. When every term in the difference (or differential) equation involves the dependent variable X(t), then the equation is called homogeneous. If additional terms, such as F(t) in Equation (8) appear, the equation is termed non-homogeneous and leads to a different solution.

In solving problems described by equations like (6), (7) or (8), we must exercise caution that the constraints on the level itself are not violated. We have already mentioned a possible constraint in the previous example: maximum (saturated) level in the number of automobiles that can be accomodated in Manhattan. Other problems will have similar constraints on the system levels, such as the maximum amount of liquid that a tank can hold, or the maximum number of units of inventory that can be stored in a warehouse, or the maximum number of people who can get sick in an epidemic (equal to the total population), etc. It should be clear that any solution predicting levels higher than these constraints is not a reasonable solution, and is based on an invalid model. Or, the level has increased beyond the maximum value for which the original model was valid.

Another constraint on the problem solution is implicit in Equation (8). The F(t) given in that equation is often specified over only a limited period of time. This means that the solution we arrive at is only valid for that same period of time. For an extended period of time, where a different F(t) is applicable, we must also find a new solution for the dependent variable X(t).

Criteria, Alternatives and Tradeoffs

Before rushing into the solution of the equations of our mathematical model, some caution and discretion are advisable. We must first ask some searching questions as to whether our model is indeed appropriate for the problem at hand, particularly with regard to the three terms used in the heading of this section. Let us examine them one at a time.

When a problem is posed to us, buried in the problem statement will be a set of criteria upon which our solution will be judged. We must first satisfy ourselves that our mathematical model really answers the basic questions raised by these implied criteria. For example, if the automobile in Manhattan problem was posed to us by the Traffic Bureau, then our previous models may suffice. On the other hand, if the problem were posed to us by the Air Pollution Board of the Environmental Protection Agency, then our model is probably insufficient. A satisfactory model, for this agency, would be one which predicts the level of carbon monoxide, carbon dioxide, sulphur dioxide, etc. in the air as a result of automobile congestion. This information is probably not directly proportional to the level of automobiles X(t) in the previous model, but is more likely a function of the type of traffic, the number of cars idling, the time of day, the weather, etc. Here it must be remembered that the criteria, implied or specific, upon which a predicted solution will be evaluated, should play a fundamental role in formulating the mathematical model from which the solution emerges. If the question is phrased poorly or incompletely, then the answer will not be meaningful.

When a problem is posed to us by our boss (or our professor), we have a natural tendency to accept "that problem statement" as the gospel and proceed immediately towards a solution. Instead, we should train ourselves to ask first whether there are any alternative approaches or formulations which are more desirable once all the criteria for evaluating the proposed solution are taken into account. To exemplify this point, assume that the Federal Aviation Agency has asked us to make a feasibility study of various proposed sites for a new major jetport to service the New York City Metropolitan area. Implicit in this job assignment are the evaluative criteria of reducing air traffic congestion at the existing jetports, as well as of controlling air and noise pollution and ground traffic congestion at these existing sites. After some amount of contemplation, particularly when looking at the presumed criteria, a completely alternative approach to a solution might come to mind. For example, we might propose that we first study the feasibility of restructuring the

airline schedules by executive fiat . . . with a sweetener added by restructuring the fare schedule as well (e.g., travelling at most popular times costs the most, and travelling at off hours yields a travel bargain). As still a different approach, we might propose studying the feasibility of creating not a major jetport, but a minor STOL (Short Take-Off and Landing) airport for short haul traffic in order to relieve the pressure from the major jetports. As still a further alternative, we might propose means of winning some of the short haul air travelers back to bus and rail transport by making the latter more attractive. For a proper study that tackles the basic problem, namely the congestion of the present jetports we must exhaust all possibilities that we can think of before commiting ourselves to a singular plan of action to be embodied by our problem formulation and mathematical model. This should be part of our professional responsibility.

In considering various alternatives, we are often faced with the situation that each one has "mixed blessings". It is usually difficult to identify a single alternative approach as being the "best". First, we must define what we mean by the "best", and second, we will usually find that alternative approaches have "tradeoffs" and that the "best" is a rather complex idea which can only be partially met. An example of a tradeoff situation is the structured fare alternative mentioned in the previous paragraph. Under what circumstances would a traveler opt to pay a premium to travel at high density travel times ? And under what circumstances would he trade off a desirable departure time for a lower fare? Obviously this depends upon the traveler's specific circumstances. This applies to most tradeoff situations; the alternatives must be carefully compared in the light of criteria for evaluating the problem solution, and a decision made as to which approach is best in balance. The necessity of evaluating criteria, alternatives and tradeoffs before formulating a specific model can not be overemphasized. Many of the technological problems bedeviling our society arose because nobody bothered to ask the right questions first.

Mathematical Models and Causality

In the previous chapter we introduced the concept of integral (accumulation) causality. Let us now examine how this relates to mathematical modeling by some additional examples.

An observation should be made at this point. All out mathematical models derived sofar (see Equations (6), (7) and (8)) are difference or differential equations which at first glance seem to have little to do with integral (or accumulated) causality. However, we will find that in the solving of these mathematical models we must go through an accumulation (for numerical) or integration (for analytical) approach to arrive at a solution. Put another way, in each mathematical model the rate of change of some level variable will be specified as an input, and the system output will come about from an accumulation or integration of that rate.

Let us consider two specific real world situations to clarify this concept. First, take the simple problem of a warehouse inventory level as it builds up due to the manufacturing rate. Let us assume that there are no current sales or pilferage to account for outflow of inventory.

We first shall try to prove that the inventory level cannot be an input by an illustration. If the level L (say, of T. V. sets) itself were truly an independent input, then we should be able to arbitrarily force it to respond in any fashion. Let us, for argument's sake, assume that at some time T we are going to instantaneously change the level from one constant value L_1 to another L_2 as shown in Figure 8. The rate of manfacture of T. V. sets $\dot{L}$ (in T. V. sets per day), had to be zero up to time T to account for the level remaining constant at L_1. Similarly, $\dot{L}$ would have to be zero for all time greater than T. But what must the rate of manufacture be in the zero time interval during which the level has jumped from L_1 to L_2? Obviously this would require an infinite rate of manufacture (infinity T. V. sets manufactured per day) which must be impossible. There can be no "zero time" during which the level jumps and we must have finite time intervals for the sharp level changes as shown in Figure 8 to allow for finite manufacturing rates. But if this is the case, we are placing a restriction on the level as an arbitrary variable and therefore it can not logically be viewed as an independent input.

On the other hand, it is not difficult to envision the possibility of step changes in the rate of manufacture $\dot{L}$, by, for example, suddenly hiring additional workers in the manufacturing plant. The plot of $\dot{L}$ versus time would be as shown in Figure 9-(a), where the manufacturing rate has gone up from $\dot{L}_a$ to $\dot{L}_b$ at time T. In Figure 9-(b) is shown the resulting level L versus time plot where we see a cusp in the level curve at point P occurring at time T with the slope of the level plot changing from $\dot{L}_a$ T. V. sets per day to $\dot{L}_b$ sets per

day as indicated.

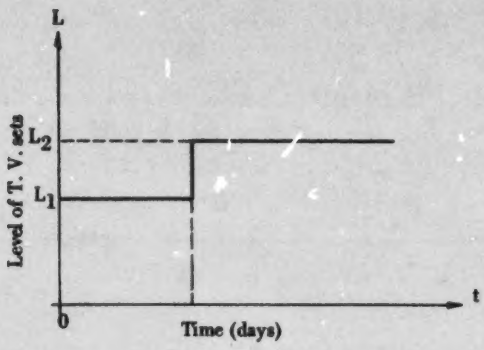


Figure 8. An impossible step change in level as a warehouse-inventory input.

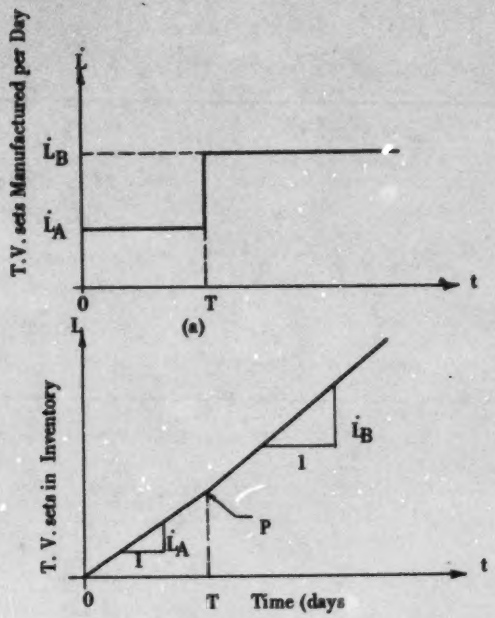


Figure 9. Manufacturing rate as an input to change output warehouse-inventory level.

We may conclude from this example that a rate of change of level can act as an input to produce output changes in that level, and not vice-versa. This, however, is not to imply that one level cannot act as an input for some other variable. For example, the warehouse inventory level L might act as an input in determining the selling price for T. V. sets recognizing supply and demand and taking cognizance of the company's current financial position. The point to remember is; there is a natural causality between a level and its rate of change.

In practical situations it is never possible to achieve a step change in the level of a variable although this can sometimes be almost achieved. In Figure 9-(a) for example, we might envision a very short but fantastic effort by many workers on overtime leading to an $\dot{L}_b \gg \dot{L}_a$ at a time T, and then the rate drops back to its normal value $\dot{L}_a$. The resulting plots of $\dot{L}$ and L versus time would appear as shown in Figure 10.

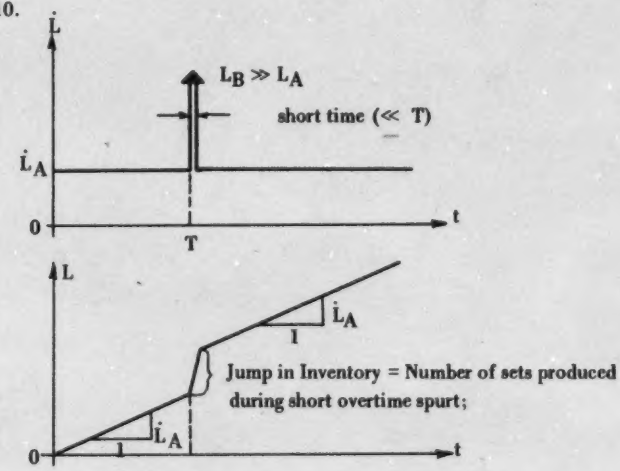


Figure 10. An apparent step change in level.

As indicated in the figure, the jump in inventory is equal to the number of sets produced during the overtime spurt. It does not actually rise vertically over zero time but has a finite slope due to the finite amount of time required to manufacture the sets. However, if this time interval is very small in comparison to the time scale of interest to the overall problem, then the jump in inventory does indeed look like a step change. Such an $\dot{L}$ input is in general called an "impulse input", and the resulting change in inventory (labelled as the "jump" in Figure 10) is equal to the area under the impulse plot (T. V. sets per day times days of overtime yields the inventory jump in T. V. sets). The area under an impulse input is then termed the strength of the input.

As our second specific real world situation, let us consider the physical problem of a single force acting upon a mass in a gravity-free environment in which there are no other forces acting on the mass except this applied force. The latter might be a rocket jet thrust as acting on a satellite in deep space. Let the level variable be the velocity of the satellite, and again let us start by assuming a step change in this level as a system input and see where this argument leads us. We will again use Figure 8, but now interpret L to be a plot of velocity (usually V) in miles per hour (mph). If we assume this change in level to be a true instantaneous step occurring in zero time (which must be possible if the velocity is to be an independent input), then we arrive at the conclusion that there must be an infinite rate of change of velocity from V_1 to V_2 at time T. But time rate of change of velocity is acceleration (mph/hour), and Newton's second law tells us that acceleration and force are proportional. Thus, assuming a step change in the velocity level calls for an impossible infinite force rocket thrust at time T.

However, the reverse situation is again possible. We can conceive of a step change in the thrust (as from $\dot{L}_a$ to $\dot{L}_b$ for example, by just turning on another jet), leading to a cusp in the velocity L versus time plot just like at point P in Figure 9-(b). The interpretation of Figure 9-(b) for this example is that the velocity had been increasing linearly with time at one thrust level, and the step change in thrust leads to the velocity increasing more rapidly with time.

We again find a situation where the rate of change of a variable is a natural input which causes changes in the level of that variable as an output, and not vice-versa. Again, this is an integral causality. We should comment in passing that there are no known practical cases of pure natural differential causality.

Linearization and Modeling

We earlier referred to the right hand side of Equation (5), $f[X(t)]$, as some as yet unspecified function of the variable X(t). In that example X(t) was the level of automobiles in Manhattan and the equation defined the rate of flow of automobiles into or out of Manhattan, $\Delta X(t)/\Delta t$. Let us now assume that this functional relationship is as given in Figure 11 (for relatively small values of X(t) that do not cause saturation of $\Delta X(t)/\Delta t$), and ask ourselves what steps we might have to take in order to solve the resulting form of Equation (5). Several alternatives are available to us.

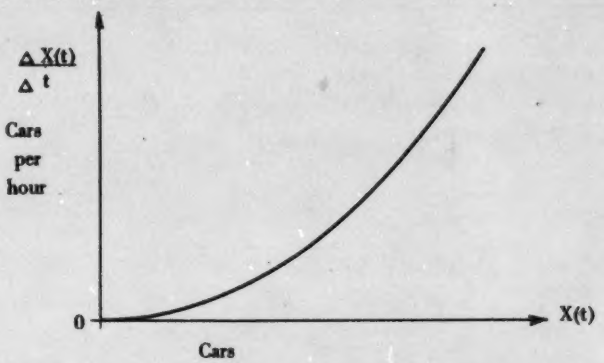


Figure 11. Typical nonlinear function for Equation (5).

Our first attempt might be to curve-fit the data of Figure 11, to arrive at an analytical expression. If for example, the curve of Figure 11 were a parabola passing through the origin, and with a unity coefficient, then Equation (5) would become

$$\Delta X(t)/\Delta t = -[X(t)]^2 \qquad (9)$$

as a difference equation, and

$$dX(t) / dt = -[X(t)]^2 \qquad (10)$$

as a differential equation. While Equation (10) is deceptively simple-looking, it is non-linear and the analytical solution of non-linear differential equations has already been referred to as a nonproductive pursuit. For an analytical approach we shall have to apply some ingenuity and reformulate the problem into a form which we can handle. Before doing so, however, let us examine Equation (9) for a moment, and apply to it the same reasoning that lead us to Equations (2) and (3) and Figure 7 in the previous chapter (i.e., a numerical solution).

Let $\Delta X(t) = [X_i(t) - X_{i-1}(t)]$, and the computing interval $\Delta t=[t_i-t_{i-1}]$ where t_i is the current computing instant and t_{i-1} is the previous one. We now find that Equation (9) becomes,

$$X_i(t) = X_{i-1}(t) - \Delta t \cdot [X_{i-1}(t)]^2 \ . \qquad (11)$$

We note that everything on the right hand side of Equation (11) is available from the previous computation, so that we are now able to compute $X_i(t)$. During the computing interval, we assume that the level $X_{i-1}(t)$ remains constant, much akin to the staircase approximation in the top half of Figure 7. The smaller the computation interval Δt, the better the approximation and the more accurate the numerical solution. Also, note that values of X(t) are known only at the computing instants, which explains why the bottome half of Figure 7 is presented as a series of unconnected dots. Regarding the choice of a computing interval, there is usually a trade-off between the accuracy sought and the allowable expense of the solution. One pragmatic approach is to start with a large value Δt, and then redo the computation with a time interval one-half that size. This procedure is repeated until the difference of the computed response between any two runs is within acceptable boundaries. It is, however, possible in a general case that for Δt far too large, no numerical solution may be found since the problem solution may diverge.

The non-linearity (squared term on the right of Equation (11)) offers us no problem, particularly if we have a computer to do our numerical work. As a matter of fact, even if the curve fitting were unsuccessful we would still be able to numerically solve this problem by utilizing a computer's memory to store a tabular form of Figure 11.

To start the computation off at i=1, we merely supply the initial value of the level X, denoted X_0. However, this initial value aspect of the problem creates a problem regarding the numerical solution. For a non-linear problem, the solution is only good for the specific initial value used. Furthermore, the numerical solution must obviously be repeated for even slightly different non-linear relations. The point to note is that in non-linear situations a numerical solution offers us no generalities upon which we can base a design approach for our problem solution.

To develop a more general analytical approach to non-linear problems, we next consider an alternative to numerical solution. Here we will take a different view of the non-linear relationship. Thus, in Figure 12, we draw a numerical representation of Equation (10) and focus our attention on the behavior of the relationship in the vicinity of a nominal operating level of the variable X (taken as $X'(t)=3$ in Figure 12). At that operating level we note that $dX(t) / dt = \dot{X}(t)$, also takes on a nominal operating value of $\dot{X}'(t)=9$ in Figure 12. The values $X'=3$, and $\dot{X}'=9$ define the point P in Figure 12. At that point we assume that the actual non-linear curved relationship is replaced by the straight line linear relationship represented by the tangent (of slope K) to the curve at point P. Making this replacement, Equation (10) becomes "linearized" as

$$dX(t) / dt = -K \cdot X(t) \ . \qquad (12)$$

We are now back to the linear differential equation. like Equation (7), for which we have promised an analytical solution in the next chapter. The question remains as to how valid (in terms of reality) that solution will be for our linearized equation, and whether its range of application is useful enough.

For the specific problem selected, a careful study of Figure 12 will give us at least a qualitative answer to this question. At X=2, the exact non-linear relation predicts an $\dot{X}$ of 4. But using the tangent approximation of the linearization predicts an $\dot{X}$ of 3, at X=2. Thus an error of δ = 1,000 cars per hour has been introduced at X=2. For values of X still further away from point P, the error introduced will be still larger, and for values of X closer to point P the error will be appreciably smaller, becoming zero at point P. Since no two non-linear relations are the same, we can make no general statement as to the size of the error introduced by linearization as a function of the departure from the linearization point. For each problem considered, we must go through a seperate similar analy-

sis, and remember to constrain our solution to a range of the variable X over which acceptable accuracy is provided by our linearization. In the present case we would probably limit our linearization analysis for the range of the variable X between about 2.5 and 3.5 .

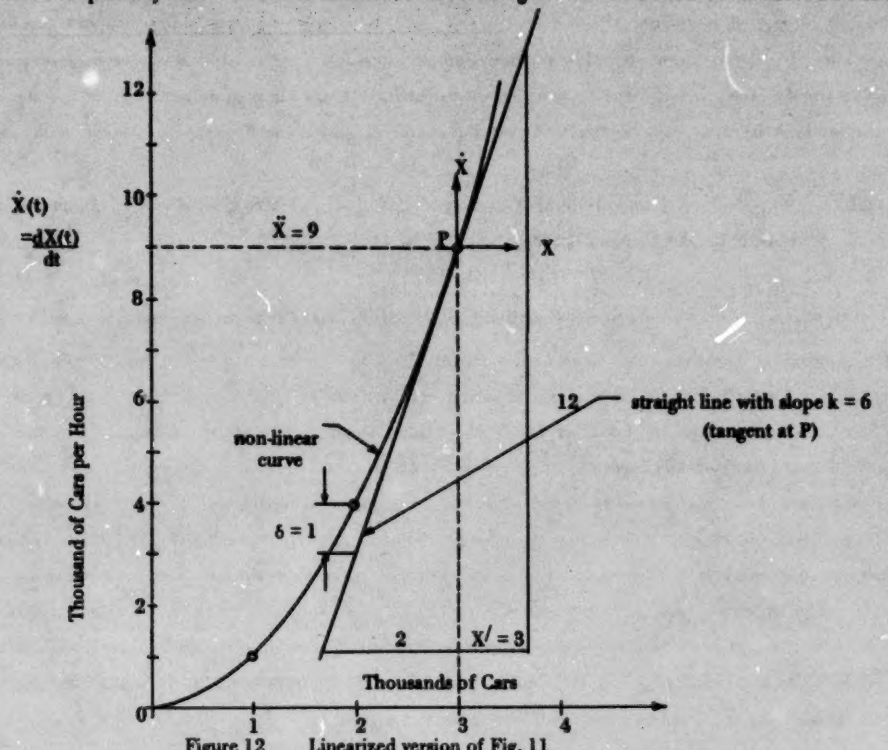


Figure 12 Linearized version of Fig. 11

In the case when a solution is sought over a wide range of the variable X(t), the tangent approximation may be repeated several times in succession. Figure 13, shows such an "envelope" of tangent approximations -- to yield a piecewise linearization of the nonlinear function. Note that even though the tangent approximations may be taken at equal intervals along the curve -- x_1, x_2, and x_3 in Fig. 13 -- the effective range of applicability of each tangent varies along the curve. By using more tangents at shorter intervals the resulting piecewise linearization will be closer to the original function with smaller errors introduced.

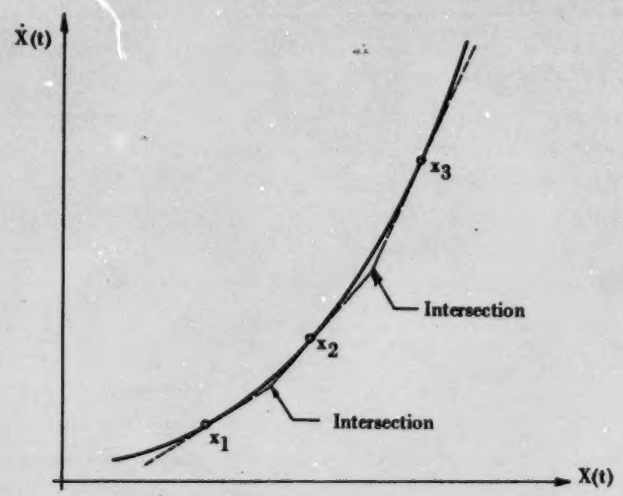


Figure 13 Succession of local tangent approximations.

This tangent approximation is the counterpart of the "staircase" approximation used in an integral representation for the area under the curve; here successive slopes are taken locally to approximate the overall behavior of the curve by local linear segments. It may also be noted that this approximation becomes poor if the slope changes too rapidly or becomes too steep.

For those of you who are mathematically inclined we make two brief observations here to relate to work in your mathematics courses. First, it can be directly shown that replacing a non-linear curve by its tangent (at a point) is directly equivalent to replacing the non-linear relationship by its Taylor series expansion at that point, and then discarding all of the higher order terms (retaining only the first order linear term). Second, we observe that Figure 12 has introduced a transformation of variables by the equations

$$\hat{X}(t) = \hat{X}' + \hat{x}'(t)$$

and

$$X(t) = X' + x(t) \,. \tag{13}$$

The significance of the $\hat{x}(t) - x(t)$ axes at point P in Figure 12 should now be clearer. In actuality then, Equation (12) is in terms of the lower case variables (as can be shown by substituting Equation (13) into Equation (10)), but the distinction is unimportant as long as we remember that the linearization is only valid in the immediate vicinity of point P.

We end this chapter with a last modeling example concerning the survival of a particular species, the sperm whale. (Ecologists believe this to be a severely endangered species which is on the brink of extinction.) Let us first model this problem in words to clarify the thought process. Let P(t) be the total sperm whale population at any instant of time and $\dot{P}(t)$ be the rate of change of that population, positive or negative. This rate of change of P(t) will depend upon the birth rate $\dot{P}_b(t)$, the natural death rate, $\dot{P}_d(t)$, and the hunting rate, $\dot{P}_h(t)$. The units of each of these terms would be whales per year, and a functional relationship can be written in shorthand notation as

$$\dot{P}(t) = f[\, \dot{P}_b(t) \,, \dot{P}_d(t) \,, \dot{P}_h(t)] \,. \tag{14}$$

To go further with this analysis we must now obtain ecological data regarding the specific functional relations embodied in the square brackets of Equation (14). Figure 14 shows a qualitative plot of $\dot{P}_b(t)$ and $\dot{P}_d(t)$ versus P(t)* which is based on past population counts. We will return to a detailed discussion of this plot in the next chapter when we consider the dynamics of this problem, but for now we merely note that we can linearize these curves at any operating level of P(t) we wish to examine. Thus, for example, if we wish to derive a linearized model of the sperm whale population dynamics in the vicinity of population level P_{eq}, we would note that (at that point),

$$\dot{P}_b(t) = B \cdot P(t)$$

and

$$\dot{P}_d(t) = D \cdot P(t), \tag{15}$$

where B and D are the local slopes of the birth and death rate curves as shown in Figure 14. The units of both B and D are whales per whale per year. Assuming that $\dot{P}_h(t)$ is some exogenous variable (international whaling agreements for example), and introducing the linearizations of Equation (15) into Equation (14), we get

$$\dot{P}(t) = [B - D] \cdot P(t) - \dot{P}_h(t) \,. \tag{16}$$

In this equation, note that both $\dot{P}_h(t)$ and the $D \cdot P(t)$ terms were entered with negative signs since they both deplete the whale population (i.e., cause a negative rate of change of whale population). Equation (16) is a linearized mathematical model of the sperm whale survival problem which we can study to investigate allowable hunting rates, $\dot{P}_h(t)$. We will do so in the next chapter.

---

* Much of this development on sperm whales draws heavily from Introducing Systems and Control, by Professors Takahashi and Auslander of the University of California at Berkeley and the author of these notes, McGraw–Hill, 1974.

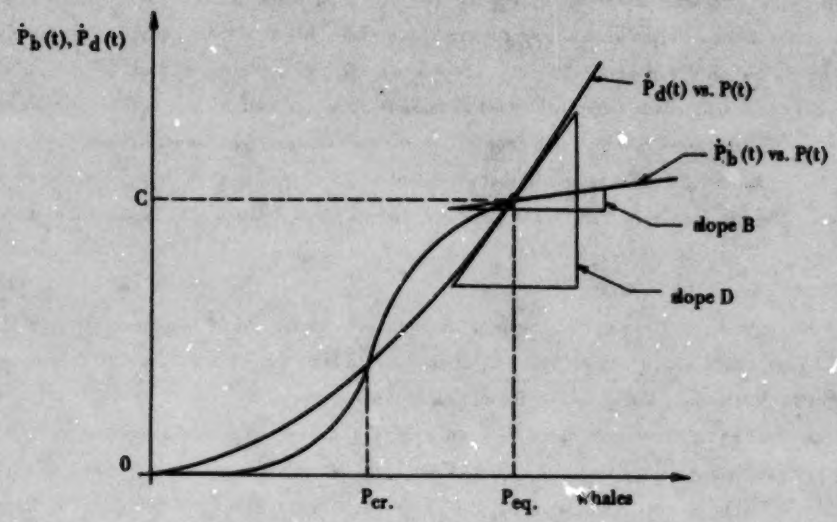


Figure 14 Whale birth ($\dot{P}_b(t)$) and death ($\dot{P}_d(t)$) rates versus whale population.

CHAPTER 3

THE SOLUTION OF FIRST ORDER EQUATIONS

Introduction

In the previous chapter we discussed the mathematical modeling of several diverse systems. In each case, the simplified model derived led to a first order difference or differential equation. Most physical problems and more realistic models usually lead to differential equations of higher order and we will develop these in the next chapter. In this chapter we will focus our attention on the solution of the first order equations derived so far to develop an understanding which will assist us in the more complex real problems. We will study three independent approaches to the solution: numerical solution, analog simulation, and mathematical analysis.

Numerical Solution

For convenience we repeat here the first order difference Equation (6) which we derived in the previous chapter to describe the rate of flow of automobiles out of Manhattan:

$$\frac{\Delta X(t)}{\Delta t} = -k \cdot X(t) \qquad (6)$$

where X is the level of automobiles in Manhattan at any time t, and k is a constant of proportionality. If t is measured in hours, k is measured in cars per hour per car, or equivalently, in inverse hours (hr^{-1}). The minus sign indicates that for this simple model the rate of flow of autos will be negative (out of Manhattan) as long as the level of cars in Manhattan is positive. Although we might like to think of negative numbers of cars in Manhattan, we will constrain our solution to deal with only positive levels of autos in Manhattan.

In the previous two chapters we have introduced the concept of a computing interval, Δt. We first discussed this in connection with Eq. (3) in Chapter 1, and then in connection with Eq. (11) in Chapter 2. Let us now utilize this concept to accomplish some actual computations.

The computing interval will be an arbitrary (for the moment) time interval between the instants t_{i-1} and t_i at which the corresponding values of the level X_{i-1} and X_i occur. During this time interval, we assume, for purposes of computation, that the rate $(\Delta X/\Delta t)$ remains constant. (This is the analytical equivalent of the staircase approximation to the rate curve shown in Figure 7.) By inspection of Eq. (6) we know that this cannot be quite true, so our computation procedure will have an inherent error built into it. We will later see that the smaller the computing interval Δt we choose, the smaller will be the error we introduce . . . . up to the limits of reasonable computing machine costs and accuracy.

In Eq. (6), let ΔX be replaced by $[X_i - X_{i-1}]$, and let the value of X on the right hand side of the equation be represented by X_{i-1}. This implies that the rate $\Delta X/\Delta t$ is based upon the last computed value of X, namely X_{i-1}, and that the rate remains constant until the next computed value of X, namely X_i, is available. With this approximation in mind, let us rearrange Eq. (6) by multiplying through by the time interval Δt and transposing terms to obtain,

$$X_i = X_{i-1} - (k \cdot \Delta t) X_{i-1} = [1 - (k \cdot \Delta t)] X_{i-1} \ . \qquad (17)$$

To start this computational procedure at i = 1, we must have available for the right hand side of Eq. (17) the initial value X_0 of the level X at t = 0.

To better comprehend this procedure, let us try to work an example through with some typical numbers. We will do this by

hand computation, although the digital computing algorithms should be fairly obvious. Let the initial level (X_0) of automobiles in Manhattan at t = 0 be 100,000 cars and let k = 0.5 hrs^{-1}. Later we will study the effect of different values of k. A value of k = 0.5 hrs^{-1} (or a 1/k of 2 hours) indicates from Eq. (6) that if the rate of $\Delta X/\Delta t$ were to continue unchanged, it would take two hours for Manhattan to empty itself of automobiles (not far off the average rush hour jam!). Of course, our problem is more complicated since the rate keeps adjusting itself to the new level as given by Eq. (6).

To show this, let us compute Eq. (17) in tabular form for several different values of Δt. The first column of Table 1 gives the elapsed time in hours to the X_i entered on that row. The next two columns tabulate X_{i-1} and $X_i = (1 - k \cdot \Delta t) \cdot X_{i-1}$ which is Eq. (17). Each horizontal entry corresponds to an increment of one in the index i. This is done for a value of Δt chosen as one hour. The procedure is repeated in the next two columns with a Δt of 0.5 hours, and then finally for a Δt of 0.25 hours. To save some writing, X has been tabulated in thousands of automobiles. Accordingly, the first entry for X_{i-1} at i = 1 is 100 (corresponding to 100,000 cars) for each Δt run. The elapsed time between any two horizontal entries of the table is equal to the Δt setting at the top of the run.

TABLE 1

($k = 0.5\ hours^{-1}$)	$\Delta t = 1$ hour $(1 - k \cdot \Delta t) = 0.5$		$\Delta t = 0.5$ hours $(1 - k \cdot \Delta t) = 0.75$		$\Delta t = 0.25$ hours $(1 - k \cdot \Delta t) = 0.875$	
Elapsed time to X_i (hours)	X_{i-1}	X_i	X_{i-1}	X_i	X_{i-1}	X_i
					X_0=100	87.5
			X_0=100	75	87.5	76.5
					76.5	67.0
1	X_0=100	50	75	56.2	67.0	58.6
					58.6	51.4
			56.2	42.2	51.4	45.0
					45.0	39.3
2	50	25	42.2	31.7	39.3	34.4
					34.4	30.1
			31.7	23.8	30.1	26.3
					26.3	23.0
3	25	12.5	23.8	17.85	23.0	20.2
					20.2	17.6
			17.85	13.4	17.6	15.4
					15.4	13.5
4	12.5	6.25	13.4	10.1	13.5	11.8
					11.8	10.3
			10.1	7.55	10.3	9.05
					9.05	7.92
5	6.25	3.12	7.55	5.66	7.92	6.93
					6.93	6.06
			5.66	4.25	6.06	5.30
					5.30	4.64
6	3.12	1.56	4.25	3.19	4.64	4.06
7	1.56	.78	Value in right hand column for each Δt equals (left hand column)$\cdot(1 - k \cdot \Delta t)$. This value is then transferred to next line in left hand column.			
8	.78	.39				

The computed response for each computational interval is plotted in Figure 15. In addition, for comparative purposes, the exact solution to Eq. (6) is also plotted in the figure. This equals $100e^{-(t/2)}$, as will be discussed in detail at the end of this chapter.

Several observations may be made from this first example. As the computational interval gets smaller, more computational

steps are obviously needed to cover the same total elapsed time. But, by the same token, smaller computational intervals bring us closer to the exact solution —— at least in the case of this example. We will find this to be uniformly true, and this becomes one of the conerstones of all digital calculations; accuracy tends to increase as the stepping interval decreases. As a matter of fact, in more complex non-linear problems, it is often impossible to obtain even an approximate solution unless the stepping interval is made quite small.

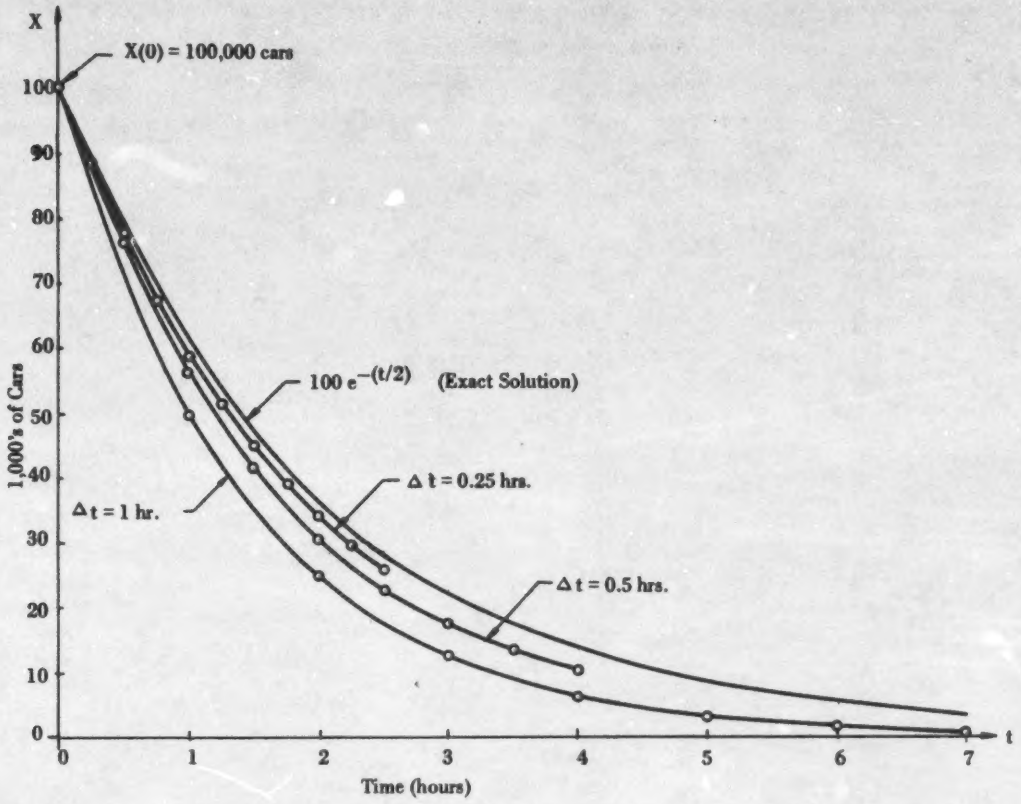


Figure 15 Response curves for various computational intervals. Shown for comparison is the exact analytical solution, $100e^{-(t/2)}$.

In practice, this would have to be balanced against the cost of computing time, which could easily become prohibitive. An extremely small computing interval is unnecessary in many cases – we are interested in accuracy only insofar as it assists the outcome of our basic problem study. Pages and pages of computer print-out with fractions of a car computed at every second would probably confuse more than help. Rather, we may wish to settle for engineering accuracy; that is, slide-rule accuracy (about ± 1%) or even ± 10% in some cases where ignorance about the basic constants would make a more accurate calculation not truly meaningful. Put another way, we are after a computation which when plotted to our scale of interest does not differ markedly from the exact response. Clearly, the Δt curve of 0.25 hours in Fig. 15 satisfies this criterion.

Fortunately, there is a rational basis upon which to proceed in the choice of a computational interval. Using estimating techniques that are yet to be discussed, we make a best guess at what the computational interval should be and then proceed. We then halve the interval and repeat the computation. If our response does not change by more than the acceptable limit (say 10% in the present linear case), then we are done. If not, we compute once more with a still smaller computational interval – one quarter of the first guess. For a system with a bounded response (one that does not become infinity) this procedure will converge.

Let us next investigate the effect of changing the value of k in Eq. (6) upon the solution to that equation. First, let us apply some qualitative considerations: as the k value gets larger, the corresponding larger rate of change of $\Delta X(t)/\Delta t$ implies that the initial level X(0) will be depleted faster. Accordingly, the solution plot of Fig. 15 should approach its equilibrium value of zero more rapidly. Conversely, a smaller k value in Eq. (6) should result in an X(t) versus t plot that takes longer to reach equilibrium.

To show that the above reasoning is correct, we recompute in Eq. (17) for three different values of k: 0.25, 0.50, and 1.0 hr^{-1}. All three computations are carried out with the same computing interval of Δt =0.50 hours. A rule of thumb for choosing the computational interval is that it should be no larger than one half of the (1/k) value. The Δt value of 0.50 hours is seen to satisfy this rule for all three k values considered here. Table 2 lists the computations, and Figure 16 shows a plot of the results.

TABLE 2
(Δt = 0.5 hours)

		k = 1.0 $hours^{-1}$ $(1-k\cdot\Delta t)=0.5$		k = 0.50 $hours^{-1}$ $(1-k\cdot\Delta t)=0.75$		k = 0.25 $hours^{-1}$ $(1-k\cdot\Delta t)=0.875$	
i	Elapsed time to X_i (hours)	X_{i-1}	X_i	X_{i-1}	X_i	X_{i-1}	X_i
1		$X_0 = 100$	50	$X_0 = 100$	75	$X_0 = 100$	87.5
2	1	50	25	75	56.2	87.5	76.5
3		25	12.5	56.2	42.2	76.5	67.0
4	2	12.5	6.25	42.2	31.7	67.0	58.6
5		6.25	3.12	31.7	23.8	58.6	51.4
6	3	3.12	1.56	23.8	17.85	51.4	45.0
7		1.56	.78	17.85	13.4	45.0	39.3
8	4	.78	.39	13.4	10.1	39.3	34.4
9		.39	.19	10.1	7.55	34.4	30.1
10	5	.19	negl.	7.55	5.66	30.1	26.4
11		negl.		5.66	4.25	26.4	23.1
12	6			4.25	3.18	23.1	20.2
13				3.18	2.38	20.2	17.7
14	7			2.38	1.79	17.7	15.5
15				1.79	1.34	15.5	13.6
16	8			1.34	1.01	13.6	11.9
17				1.01	.75	11.9	10.4
18	9			.75	.56	10.4	9.1
19				.56	.42	9.1	7.95
20	10			.42	.32	7.95	6.95
21				.32	negl.	6.95	6.08
22	11			negl.		6.08	5.32
23						5.32	4.66
24	12					4.66	4.08

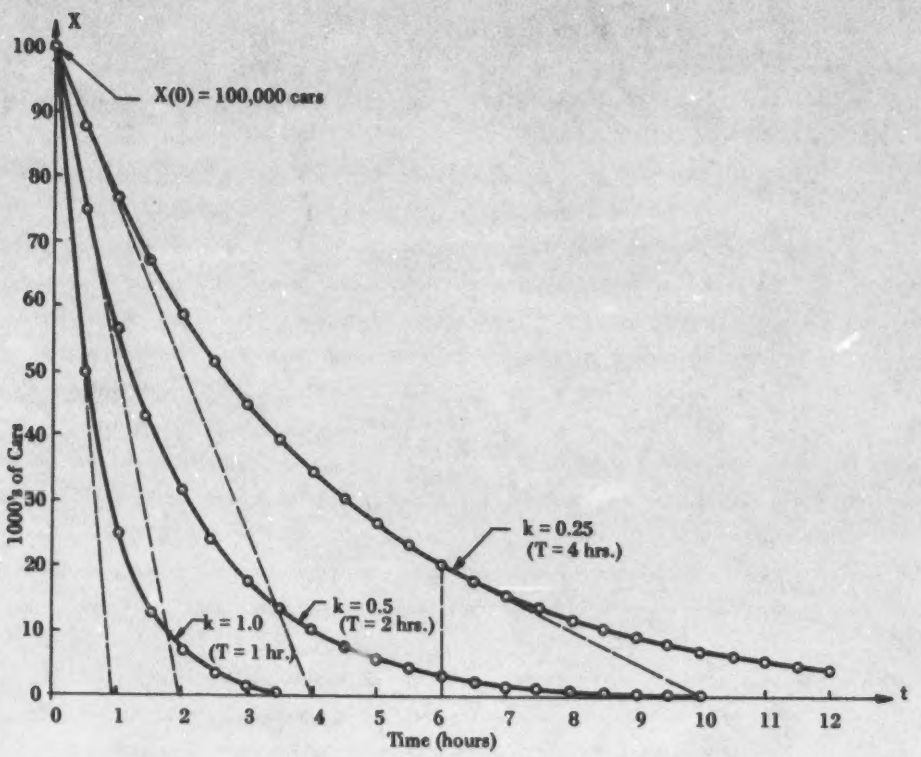


Figure 16. Response curves for various k values in Eq. (6). Dotted lines indicate tangents to the curve and are discussed in the text. (T = 1/k)

From Fig. 16, we see that our qualitative reasoning was correct. We further note that all three response plots asymptotically approach the equilibrium level of zero cars in an exponential fashion. While we will reserve the detailed discussion of the characteristics of an exponential plot until the last section of this chapter, we can make some semi-quantitative observations at this point. We note in Fig. 16 that a tangent drawn to any of the three curves at t=0 (the initial slope, drawn dotted in the figure) will intersect the horizontal time axis at a value given by (1/k). Furthermore, the tangent to any curve drawn at any time will intersect the horizontal axis (1/k) time units beyond that time, as indicated for example at an elapsed time of 6 hours on the k = 0.25 curve.

The significance of the tangents drawn in at t=0 in Fig. 16 may be deduced by consideration of Eq. (6). A computation at t=0, based on the initial level of cars X(0), gives the initial rate of change of the level of cars $\Delta X(0)/\Delta t$. But we saw in the previous chapter that this rate of change of the level is the slope of the level curve, in this case at t=0. Now if this rate of change of the level of cars were to remain constant, then we could easily compute the time needed to totally deplete the initial level of cars. For k=1.0, 0.5, and 0.25, the computed initial rates are 100,000, 50,000, and 25,000 cars per hour respectively; and if these rates were to remain constant (as signified by the dotted lines), the time for the initial level of 100,000 cars to be totally depleted would be one, two and four hours respectively. However, the rates do not remain constant, and the actual cars versus time plots follow the solid line curves rather than the imaginary dotted line trajectories. The entire discussion of this paragraph is equally applicable at values of time other than zero. The tangent drawn at 6 hours on the k=0.25 curve indicates that for this case X(6) would be depleted after 10 hours if the slope at six hours were continued unchanged. This is equivalent to defining a new initial value problem starting at six hours.

At this point we should be able to extend the discussion regarding the solution dynamics for systems which are described by a difference (or differential) equation like Eq. (6). To do so we will replace k (which has units of hours^{-1}) in Eq. (6) with 1/T where T thus has units of hours. The resulting equation is,

$$\frac{\Delta X(t)}{\Delta t} = -X(t)/T \ . \qquad (18)$$

The quantity T is termed the system time constant. One definition of T is the time it would take for the initial level of the dependent variable X(0) to be depleted, if the initial rate of change of that variable $\Delta X(0)/\Delta t$ were continued unchanged. At this point we can conclude that systems with small time constants reach equilibrium faster than do systems with relatively long time constants. If one of the objectives in a given system study is to have transients decay more rapidly, we must seek ways of creating smaller system time constants.

Not all first order systems reach their equilibrium at zero, and in fact some first order systems never do come to equilibrium. To exemplify such situations, we will quantitatively study the sperm whale survival problem modeled at the end of the last chapter. Consider Fig. 14 which is a plot of both the whale birth rate $\dot{P}_b(t)$ and death rate $\dot{P}_d(t)$ versus the instantaneous whale population P(t). From this plot considerable information is available concerning the dynamics of this ecological system.

If the current whale population falls between the $P_{cr.}$ and $P_{eq.}$ values indicated on Fig. 14, then the birth rate is greater than the death rate as shown and (in the absence of any hunting) the whale population will increase. If the population were to increase beyond $P_{eq.}$, then the situation reverses with the death rate predominating over the birth rate as shown in the figure, so the whale population would naturally tend to stabilize at the non-zero equilibrium value of $P_{eq.}$. If the whale population were to fall below the critical value $P_{cr.}$, then the death rate would be higher than the birth rate and the whale population would be driven to a zero equilibrium level resulting in extinction. At this time, the sperm whale is close to its critical population level. Other species have fallen below their critical levels and have disappeared, or been maintained through stringent governmental actions* as in the case of the whooping crane and the buffalo.

The actual whale population problem is somewhat more complicated when hunting rates as well as birth and death rates are considered. As a final numerical solution example, let us consider the case when we have applied the linearization technique of Fig. 12 in the vicinity of $P_{eq.}$ in Fig. 14 and we simultaneously consider a specific constant hunting rate, $\dot{P}_h(t)$. The linearization means that we are now considering a new variable, which we will call p(t), with origin at $P_{eq.}$, such that

$$p(t) = P(t) - P_{eq.} \ . \qquad (19)$$

In terms of this new variable, the linearized birth and death rate expression of Eq. (15) now become,

$$\dot{P}_b(t) = B \cdot p(t) + C$$

and

$$\dot{P}_d(t) = D \cdot p(t) + C \qquad (20)$$

where C is the ordinate height on Fig. 14 to the intersection of the birth and death rate curves at $P = P_{eq.}$. The total rate of change of whale population is given by,

$$\dot{P}(t) = \dot{P}_b(t) - \dot{P}_d(t) - \dot{P}_h(t). \qquad (21)$$

Substituting Eq. (19) into Eq. (20), the result into Eq. (21) and slightly rearranging terms, we obtain:

$$\dot{P}(t) = (B-D) \cdot P(t) \ - (B-D) \cdot P_{eq.} - \dot{P}_h(t). \qquad (22)$$

Comparing this last equation to the form of Eq. (6) two immediate observations can be made: first, the last two terms on the right hand side of Eq. (22) have not appeared before, and second, the term (B–D) plays the same role as k did earlier. Thus, the inverse of (B–D) is the system time constant and it does indeed have the units of time (years), since B and D are both measured in units of whales per year per whale.

To numerically solve Eq. (22), we will replace $\dot{P}(t)$ on the left hand side of the equation by $\Delta P(t)/\Delta t$ to form a difference equation, multiplying through by Δt, and replacing $\Delta P(t)$ by $(P_i - P_{i-1})$. Doing so, and replacing P(t) on the right hand side of the

---

* Such actions reduce the $P_{cr.}$ level because normal balances are redressed – i.e., decreased normal loss due to moderate hunting, increased feeding grounds, etc. Basically, the constraints on the problem have been changed! Since the whale has few natural (non-human) enemies, $P_{cr.}$ is already very low and excessive hunting can cause a species to disappear precipitously.

equation by P_{i-1}, we get

$$P_i = P_{i-1} + [\Delta t \cdot (B-D) \cdot P_{i-1}] - [\Delta t \cdot (B-D) \cdot P_{eq.}] - [\Delta t \cdot \dot{P}_h(t)] . \qquad (23)$$

This equation is the algorithm for numerically solving the linearized whale population problem. To go through a sample numerical solution, using convenient values which are not necessarily realistic, assume the following values:

$B = 0.5$ whales born per whale per year,

$D = 1.5$ whale deaths per whale per year,

$P_{eq.} = 2{,}000$ whales,

$\dot{P}_h(t) =$ a constant 500 whales killed by hunting per year,

and $P(0) =$ initial whale population at 2,500 whales.

Substituting these numerical values into Eq. (23) yields

$$P_i = P_{i-1} - \Delta t \cdot [P_{i-1} - 1{,}500], \qquad (24)$$

Comparing this equation with the previous computing algorithm of Eq. (17), it is seen that the k in Eq. (17) has now been replaced with the unity numerical value of (B–D). More important, however, is the appearance of the constant term of 1,500 on the right hand side of Eq. (24), which arose from the linearization process and the imposition of a whale hunting rate. (This latter change would be like superimposing a constant in- or out-flow of a given number of automobiles per hour on the previous traffic problem.) It is this constant term which results in a non-zero equilibrium value.

Since (B–D), and therefore the system time constant, has a value of unity, we will choose a computational interval of ¼ of a year, which is less than the maximum computing interval of one half of the system time constant. Table 3 lists the computations and the results are plotted in Fig. 17.

In Fig. 17 an exponential dynamic response is again apparent, but with a non-zero equilibrium level of 1,500 whales. This equilibrium level is a combination of the equilibrium level from Fig. 14 assumed to be $P_{eq.} = 2{,}000$ whales and the assumed annual hunting rate of 500 whales. Clearly, this equilibrium level will shift as the annual allowable hunting changes. We obviously should never allow this equilibrium to fall below the $P_{cr.}$ value of Fig. 14 by permitting too high an annual catch, otherwise the species will die out.

TABLE 3

$\Delta t = (¼)$ year

i	Time Elapsed to P_i	P_{i-1}	$[P_{i-1} - 1{,}500]$	$\Delta t \cdot [P_{i-1} - 1{,}500]$	$P_i = P_{i-1} - \Delta t \cdot [P_{i-1} - 1{,}500]$
1		P_0=2,500	1,000	250	2,250
2		2,250	750	187.5	2,062.5
3		2,062.5	562.5	140.5	1,922
4	1 year	1,922	422	105.5	1,816.5
5		1,816.5	316.5	79.1	1,737.4
6		1,737.4	237.4	59.4	1,678
7		1,678	178	44.5	1,633.5
8	2 years	1,633.5	133.5	33.4	1,600.1
9		1,600.1	100.1	25.	1,575.1
10		1,575.1	75.1	18.8	1,556.3
11		1,556.3	56.3	14.1	1,542.2
12	3 years	1,542.2	42.2	10.5	1,531.7
13		1,531.7	31.7	8	1,523.7
14		1,523.7	23.7	6	1,517.7
15		1,517.7	17.7	4.4	1,513.3
16	4 years	1,513.3	13.3	3.3	1,510.

If P(0) were changed from 2,500 to 1,000 whales (assuming that this is still greater than $P_{cr.}$) and all of the other numerical values were unchanged, then the resulting plot of whales versus time would be an exponential rise which asymptotically approaches $P_{eq.}$ from below. Furthermore, if P(0) happened to be at the equilibrium level $P_{eq.}$ at the start of the problem, then Eq. (24) indicates that the whale population would remain constant for all time, and there would be no dynamics in the problem.

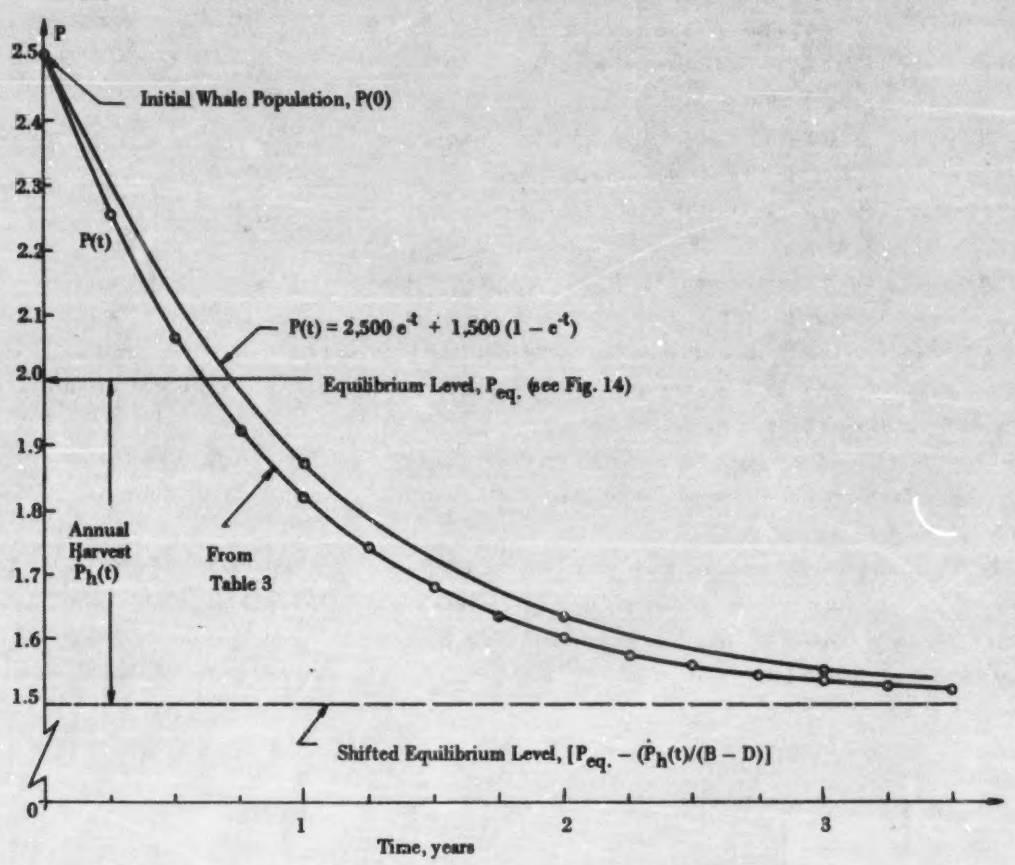


Fig. 17 Whale population versus time, from Table 3. See the end of this chapter for development of the theoretical exact curve.

Finally, a warning about the linearization process should be repeated. The numerical values for B and D that were substituted into Eq. (20) were evaluated by taking the slope values on Fig. 14 at the $P_{eq.}$ level! As soon as the current instantaneous value of P(t) differs appreciably from $P_{eq.}$, new values of B and D must be found from Fig. 14. For purposes of this solution it was assumed that B and D do not change appreciably over the range of values of P(t) plotted in Fig. 17, which may or may not correspond to the true situation.

Analog Computer Solution

Having laid the groundwork in the previous section for digital computer solution of dynamic problems, it is appropriate to next discuss another kind of computer solution for the same kind of problem. At the end of Chapter 1, brief reference was made to analog simulation and the integrating operational amplifier of an analog computer. Let us now explore this topic without going into the electrical engineering details, but rather with the aim of understanding how to use analog techniques. An operational amplifier (op-amp) is the basic building block of an analog computer. It is a highly stable and accurate d-c amplifier*. Most commercial

* Footnote on next page.

analog computers use op-amps which operate in a range of zero to ten volts, while some also operate in the zero to one hundred volt range. Since currently available analog computers are predominantly ten volt systems, discussion is limited to these units.

Due to the inherent circuit characteristics of an amplifier, the output voltage of the op-amp will be the negative of the input, and may conventionally be arranged to be 1/10, 1, or 10 times the input voltage depending on the internal circuits and the external connections. We will not refer to specific wiring or patching diagrams in these notes since the various computer companies arrange their patching connections differently. Instead, we will draw generalized patching diagrams from which the wiring connections may be deduced for each particular machine. Thus, in Fig. 18-a we introduce the standard notation for an op-amp; the triangle drawn as shown. The line coming out of the triangle to the right is the output voltage, labeled e_o. The various lines coming into the triangle from the left are input voltages which the op-amp will internally sum to create the output voltage. The number entered in the triangle at the right end of the input voltage lines refers to the gain or amplification to which the input was connected. The situation shown in Fig. 18-a is typical and represents an output voltage given by

$$e_o = -e_{i1} - e_{i2} - 0.1e_{i3} - 10.0\, e_{i4} \quad .$$

The various input voltages on the right hand side of this last equation may be supplied from constant or time varying voltage sources, or may be tapped off from other points in the analog computer circuit.

The second basic building block of an analog computer is a specially wired op-amp arranged such that the output voltage is the negative integral (accumulation) of the input voltage(s). It is called an integrating op-amp, and its standard circuit depiction is as shown in Fig. 18-b. Again, the output voltage is the negative sum of the integrals of the input voltages, and the input voltages may generally be wired with gains of ten, one, or one-tenth as shown in the figure.

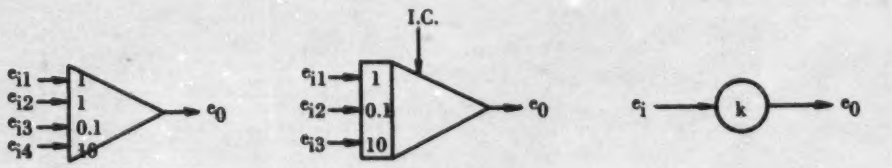
As we have seen in all of the computations we have so far done, we must always start with an initial value corresponding to time zero to get the solution going. The same is true in the case of the analog computer simulations. The integrating op-amp is internally arranged so that an arbitrarily set initial value of the output voltage is impressed on the output line of the op-amp until the "Operate" button is pushed. At that point, the initial condition voltage, I. C. in the figure, is switched out and the patched input voltages are switched in. Accordingly, the typically arranged integrating op-amp of Fig. 18-b behaves, as

$$e_o = \text{I. C.} \quad \text{for } t < 0,$$

and

$$e_o = -\int_0^t (e_{i1} + 0.1\, e_{i2} + 10.0\, e_{i3})\, dt \quad \text{for } t \geq 0.$$

To obtain a plus sign, it would be necessary to add another op-amp with unity gain (a "sign inverter") immediately following the integrating op-amp.



(a) The op-amp (b) The integrating op-amp (c) The coefficient multiplier

Figure 18 Analog Computer Elements.

In this last session, the integrating sign $\int_0^t$ is read "integral from zero to time t" and signifies that the output voltage is the accumulation of the area (voltage) under the summed input plot up to the time that the integration is stopped by pushing the "Hold" button on the machine. Some machines are arranged to integrate for a set period of time (say, one second real time) and then

*(Footnote from previous page) A "d–c" amplifier is one that can operate on only very slowly varying transients which have response times which are very large (seconds or minutes of real time) compared to "a–c" transients which change at frequencies of 60 Hertz (cps) or higher. An operational amplifier in an analog computer is an electronic unit in which input voltage signals are added, subtracted or integrated while they may at the same time be amplified.

hold the output for recording purposes and do not continue the integration until the "Operate" button is again depressed.

The last basic building block required is a coefficient multiplier. It is nothing more than a variable resistance, called a potentiometer, or pot, and is represented in patching diagrams as shown in Fig. 18-c, where

$$e_o = k \cdot e_i .$$

Note that there is no sign reversal in this case and that for each simple coefficient multiplier there may be only one input. Because of the characteristics of the potentiometer, the value of the coefficient multiplier must lie between zero and unity. Incidentally, the arrowheads in Fig. 18 are important and signify the unilateral flow direction of both information and electrical currents.

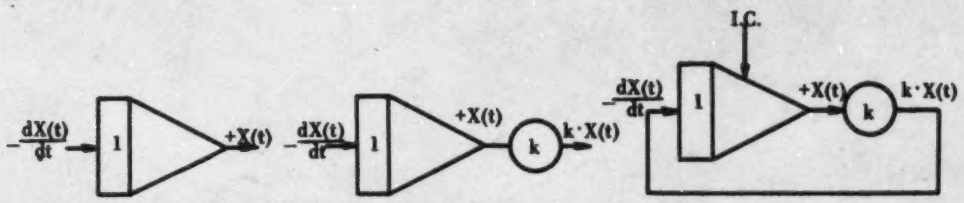
One point should be emphasized before embarking on actual sample analog simulations. The op-amps are designed to operate correctly only within the voltage limits of the machine in use. If any voltage in an analog computer circuit exceeds ten volts in a ten-volt machine, then the op-amps will no longer operate in the linear fashion described above. Usually, they will saturate and the output voltages will flatten out at ten volts and cannot go any higher no matter what the input voltages demand. This becomes a particularly sensitive problem for integrating op-amps where the output accumulates all past inputs. Some machines have warning over-load lights to indicate when any op-amp output exceeds the machine limit, but in any event caution must be exercised that the machine always operates within its specified limits.

At this point in the discussion, most texts on analog computing embark on the topic of scaling which deals with the problem of assigning equivalent voltages to problem variables such that the machine will always operate within its specified limits. We will leave the details of this topic to the many texts available on the subject, and focus our attention instead on some sample analog simulations. In the process, we will examine the scaling problem as it arises.

To start with the simple traffic problem which initiated this chapter, let us pose it in the form of the differential Eq. (7) rewritten here as

$$-\frac{dX(t)}{dt} = k \cdot X(t), \qquad (25)$$

where for convenience the minus sign is transposed to the left hand side of the equation. In analog simulation work we start by assuming that the highest derivative is available. In this case it is -dX(t)/dt and we supply that as the input to an integrating op-amp as shown in Fig. 19-a. According to the previous discussion, the output of the integrating op-amp will be the negative integral of the input or +X(t). In physical terms, accumulating the rate of change of a level brings us back to that level – as seen in detail in Chapter 1.



(a) First step (b) Second step (c) Completed patch diagram

Figure 19 Analog simulation of traffic problem.

Having produced +X(t) as a variable in Fig. 19-a, we now introduce a coefficient multiplier in the output line as shown in Fig. 19-b, to obtain k times this value. But now the output variable exactly equals what the input variable must be according to Eq. (25). Accordingly, the output is connected back to the input as shown in Fig. 19-c to complete the patching diagram for the problem. The pot setting should be adjusted to the k value desired, 0.5 in this case. If a k value of 5.0 were desired, the pot would still be set at 0.5, but the input to the op-amp would be set to a gain of 10.0 rather than 1.0.

To start the solution, an initial value for X(t) must first be supplied, and to do this an equivalence between machine voltage and problem variable must first be decided upon. Using the same data as for the earlier numerical solution, assume an initial level of 100,000 cars in Manhattan. Presuming a ten volt machine, a first try at scaling would be to let one volt represent 10,000 cars (not

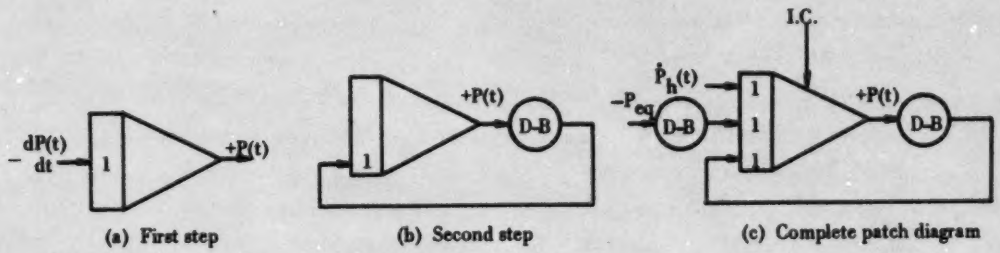
many significant figures in the results will be available here unless we have highly accurate op-amps and voltmeters accurate out to the fourth or fifth decimal place – a typical situation in analog simulation. Accuracy is a separate question. What this scaling means is that the initial voltage setting of I. C. should be 10 volts, which will work al long as the X(t) variable in the problem never exceeds 100,000 cars. This limit is known to be true from the numerical analysis. But even if the latter had not been carried out, this scaling approach might still be tried while watching the op-amp voltage to make sure that it never goes above ten volts. If it does, scaling would then have to be redone and the I. C. voltage accordingly reduced.

Having started our analog situation going, it only remains to complete a run and record the results. One second of actual machine computing time is arbitrarily set equal to one hour in the traffic problem, noting that in the process a simulated solution is produced that runs many times faster than the real problem. (It is possible as well to slow up real world problems in simulations.) To complete our simulated solution to the traffic problem we merely record the op-amp output voltages* as a function of time and translate the voltages into levels of automobiles by multiplying each voltage we wish to plot by our scaling factor, 10,000 cars/volt. The resulting plot would appear as the theoretical exact curve of Fig. 15, except that each second of machine time corresponds to one hour of real time.

As an example of a more complicated analog simulation, reconsider the whale survival problem described by Eq. (22). For convenience, multiplying both sides of that equation by minus one and replace –(B – D) by its equivalent +(D–B) so that positive coefficients remain. Thus,

$$\frac{-dP(t)}{dt} = (D - B) \cdot P(t) - (D - B) \cdot P_{eq.} + \dot{P}_{h}(t). \qquad (26)$$

To perform an analog simulation of this problem, again start by assuming the availability of the highest derivative term, in this case –dP(t)/dt as shown in Fig. 20-a. The output of the integrating op-amp would be +P(t) as discussed in the previous example.



(a) First step (b) Second step (c) Complete patch diagram

Fig. 20. Analog simulation of the whale problem.

According to the right hand side of Eq. (26), (D–B) times the variable P(t) is needed in order to construct –dP(t)/dt, so a (D–B) coefficient multiplier is inserted and that signal is fed back as shown in Fig. 20-b. But this does not yet completely specify –dP(t)/dt as can be seen by the two additional terms on the right hand side of Eq. (26). Therefore, two additional input voltages must be supplied to the left hand side of the integrating op-amp as shown in Fig. 20-c. The first is a constant voltage proportional to the annual whale hunting rate $\dot{P}_h(t)$, and the second is a voltage proportional to the negative of the whale equilibrium level $P_{eq.}$. This last input voltage must be fed through a coefficient multiplier which equals the first multiplier setting of (D–B). Now the sum of the three input voltages in Fig. 20-c is equivalent to the right hand side of Eq. (26) and thus produces the variable –dP(t)/dt... the start in this loop procedure. Analog simulation is thus seen to be similar to a bootstrap operation where is is assumed that the highest

*Depending upon the laboratory equipment available, this recording may be accomplished in various ways. Some analog computers have a built-in time which permits the solution to proceed for a set arbitrary time, and then holds the output voltages for voltmeter reading until the operate button is again depressed. Another approach would be to make a strip chart recording of the output voltage where distance along the chart corresponds to elapsed time.

derivative is available, and then we go ahead to produce that derivative.

Again, to initiate the solution, an initial voltage must be applied to the integrating op-amp, which means that some scaling relating voltage to whales must first be accomplished. Assuming the same numerical data as given just before Eq. (24) and a 10 volt machine, scaling should be such that at least the initial value of 2,500 whales is accomodated on our machine. If this is the maximum number of whales encountered in the problem, we can let one volt represent 250 whales, so that the initial voltage I. C. setting will be 10 volts. Of course, the simulation run should be checked immediately to insure that the whale level P(t) never exceeds 2,500 throughout the run or that the machine voltage never exceeds ten volts.

However, before starting a simulation run, two other voltages still have to be set in this problem. To match 10 volts corresponding to 2,500 whales, the constant annual kill of 500 whales is represented by a +2 volt input for $\dot{P}_h(t)$, and the equilibrium level of 2,000 whales is represented by a −8 volt input for $-P_{eq.}$. Assigning one whale year to each second of analog computer time, the actual simulation solution is ready to be run. To translate the voltage results into whales, each voltage to be plotted need only be multiplied by our scale factor of 250 whale per volt, and each second on the machine interpreted as one year. The results would be Figure 17.

In the next two chapters we will consider analog simulation of more complicated problems, leading to the use of several op-amps and the possibility of interconnected loops. However, before leaving the subject, let us make one more observation. If the quantity (D−B) turned out to be negative, due to a higher birth rate than death rate, what would this mean in Fig. 20-c and in the resulting solution? First, recall that a coefficient multiplier can be set with only positive coefficients, so (as noted earlier) in order to produce the required minus sign an op-amp with unity gain would have to be added in the feedback path of Fig. 20-c to act as a sign inverter. The result would be an even number (2) of op-amps in the loop of Fig. 20-c. But, what is the physical significance of B being greater than D? As long as P(t) is greater than the critical value $P_{cr.}$ of Fig. 14, and the hunting level $\dot{P}_h(t)$ is not sufficient to force P(t) below $P_{cr.}$, then the condition of B being greater than D will result in an exponentially growing whale population P(t). The op-amps, if not the oceans, will soon saturate. One heuristic conclusion we may make, which turns out to be true in general, is that any analog computer simulation containing an even number of op-amps in any loop will produce an exponentially growing solution and a simulation run that will quickly overload the machine.

Mathematical Analysis

Having considered both numerical and analog computer solution of first order difference and differential equations, we are left with one last direction to explore, which actually falls in the territory of Differential Equations. We have no intention of either summarizing or replacing a course in this field, but only to introduce and explore those characteristics of the analytical solution of differential equations which provide additional understanding to our work up to now.

Let us again reconsider the traffic problem of Eq. (7), which is now written in the form,

$$\frac{dX(t)}{dt} + k \cdot X(t) = 0. \qquad (27)$$

The zero on the right hand side of this equation signifies no external forcing function or excitation acting upon the traffic level in Manhattan, such as would occur if a specified flow rate into the borough would bring cars in while at the same time the traffic is slowing out. Such cases will be left until later; for now we will consider the solution of the equation with a zero on the right hand side, the so-called homogeneous equation case.

It is convenient to introduce a short hand notation at this point for equations of the form of Eq. (27). The derivative operation [d /dt] is replaced by the single letter D to form

$$D \cdot X(t) + k \cdot X(t) = 0\,, \qquad (28)$$

where the first term is read "D operating on X." This symbolism is accordingly called "operator notation" and at this point is only a notational convenience.

Equation (28) is next rewritten by factoring out the common term X(t) to obtain

$$(D + k) \cdot X(t) = 0, \qquad (29)$$

which is read as "D plus k operating on X." A trivial solution to Eq. (29) occurs if X(t) is always zero. If a non-trivial solution is to

apply for all time, then

$$(D + k) = 0 \qquad (30)^*$$

must be true, or

$$D = -k. \qquad (31)$$

Since a course in differential equations is not part of our present study, we have to give the mathematical solution of Eq. (29) without proof as

$$X(t) = Ce^{D \cdot t} = Ce^{-k \cdot t} \qquad (32)$$

Although students familiar with the elements of differential calculus can readily show differentiation with respect to t will again lead to Eq. (28) or (29). In this equation, e is the natural (Naperian) logarithmic base of 2.7318..., D is the root of the characteristic equation, as given by Eq. (31), t is the independent variable (time in this problem), and C is a constant of integration that must be evaluated from the initial conditions. Here we are told that at t=0, the initial level of cars in 100,000. Substituting this information into Eq. (32), we find

$$X(0) = 100{,}000 = C \cdot e^{D \cdot 0} = C \qquad (33)$$

since $e^0=1$. Note that this initial value information is a prerequisite for a solution, no matter what solution path is taken.

Substituting Eqs. (31) and (33) into Eq. (32) and recalling that k=—D was earlier replaced by 1/T, the transient solution to the homogeneous case is obtained as

$$X(t) = 100{,}000 \cdot e^{-(t/T)} \qquad (34)$$

Referring back to Fig. 15 with "thousands of cars" plotted as the ordinate variable, and where the time constant T was specified as two hours, Eq. (34) is seen to be precisely the theoretical exact curve plotted as the right-most curve.

We next complicate the problem somewhat (and make it more realistic) by specifying that simultaneous to the outflow of traffic from Manhattan, as given by Eq. (7), there will also be a constant inflow of I cars per hour. This term will appear as a positive addition on the right hand side of Eq. (7), so Eq. (27) becomes

$$\frac{dX(t)}{dt} + k \cdot X(t) = I. \qquad (35)$$

Now we no longer have a homogeneous equation to solve, and we again have to rely on mathematical techniques to identify a solution approach. These suggest that the total solution to Eq. (35) is made up of two parts: the first is the solution to the homogeneous equation and in this case it is identical to Eq. (32). This part is the so-called transient or force-free solution. The second part of the solution arises from the forcing function on the right hand side of Eq. (35) -- in this case tne constant value of I. This results in the so-called particular, steady-state, or forced portion of the solution and the total response will be given by the sum of the transient and steady-state portions of the solution.

To find the steady-state part of the solution we make use of the fact that it is known to be of the mathematical form as the forcing function (in this case a constant) as well as all of the derivatives of the forcing function (zero in this case). The task of identifying the steady-state solution is usually made easy by the inclusion in most Differential Equation text books of tables of steady-state solutions for various forcing functions. Let us assign some as yet unknown constant value X_{ss} as the steady-state solution and substitute it back into Eq. (35) to find its value. Recognizing that the derivative (rate of change) or a constant is zero, we find

$$0 + k \cdot X_{ss} = I,$$

$$\text{or} \qquad X_{ss} = I / k \quad . \qquad (36)$$

---

*Equation (30) is called the characteristic equation of the system, and is found in general by applying operator notation to the system's homogeneous differential equation. The left hand side of the equation is called the characteristic function. Solution of the characteristic equation polynomial in the operator D, such as in Eq. (31) for this simple first order case, yields the system's characteristic or eigenvalues. These roots, such as -k in this case, give us much information about the system's characteristic dynamic response, as has already been seen in our numerical solution section.

Eq. (36) is the particular or steady-state solution to Eq. (35) and the total solution will be given by

$$X(t) = C \cdot e^{D \cdot t} + I/k. \tag{37}$$

To find the constant of integration C, we again make use of our initial value X(0). Note carefully that we do not attempt to evaluate the constant of integration until the transient and steady-state solutions are added to form the total solution as in Eq. (37). Letting t = 0 in Eq. (37) and using X(0),

$$X(0) = C + I/k,$$

or

$$C = X(0) - I/k \ . \tag{38}$$

Substituting this value for C back into Eq. (37) and rearranging terms leads to

$$X(t) = X(0)e^{D \cdot t} + (I/k)[1 - e^{D \cdot t}] = X(0)e^{-kt} + (I/k)[1 - e^{-kt}] \ . \tag{39}$$

Let us examine this solution and make several observations about it. First, we note that at t = 0, X(t) reduces simply to X(0) as it must. Second, we see that all terms in Eq. (39) are dimensionally consistent: X(t) and X(0) are measured in cars, D is equivalent to k which we earlier saw to be inverse hours so that (D times t) is dimensionless as it must be for mathematical consistency and I cars per hour divided by k hours^{-1} again yields cars. Third, the first term on the right hand side of the equal sign in Eq. (39) is exactly the same as the transient solution found in Eq. (34) for the homogeneous case. Finally, we observe that the particular solution appearing as the last term on the right hand side of Eq. (39) introduces a non-zero equilibrium level to the solution, which is exactly akin to the case of the non-zero equilibrium level which arose earlier in the whale survival problem.

To convince ourselves about the nature of the particular solution portion of Eq. (39), consider some typical numbers: take k = 0.5 as before and I as 10,000 cars per hour steadily flowing into Manhattan. Letting t approach infinity, both exponential terms in Eq. (39) go to zero, and we are left with a steady-state equilibrium value of X(∞) of 10,000/0.5 or 20,000 cars. This tells us that when there are 20,000 cars in Manhattan, then the steady outflow of k·X(t) or 10,000 cars per hour will exactly balance the steady inflow rate of 10,000 cars per hour, and the level of cars in Manhattan will not change.

It should be reasonably clear that these same solution procedures applied to the non-homogeneous whale problem given by Eq. (22) will result in a solution similar in form to Eq. (39). Going through the actual steps results in

$$P(t) = P(0)e^{-(t/(D-B))} + [(D - B)P_{eq} - \dot{P}_h][1 - e^{-(t/(D-B))}]. \tag{40}$$

Substituting in the numerical values given just before Eq. (24) results in the theoretical "exact curve" plotted in Fig. 17.

In the next two chapters, we will be examining some more realistic problem models which result in second and higher order systems. While we cannot extend the work of this chapter completely to higher order systems, we will, within the scope of these notes, and as appropriate, carry on some of the reasoning and techniques introduced herein to gain a better understanding of more complex systems.

CHAPTER 4

SECOND ORDER SYSTEMS

Introduction

In most real world situations, even simplified mathematical versions of a physical model turn out to be of higher than first order. In this chapter we will study the modeling, numerical solution, analog simulation and analysis of second order systems and conclude with some comments on extensions to higher order systems. Throughout the chapter the ideas developed previously will be used as building blocks.

Second Order System Models

Imagine that our assignment is to model the dynamics of the armaments race between world powers to pinpoint the causes of a spiraling arms build-up. As a start, let us simplify the problem by considering only two competing countries without including other interested nations, and linearize the problem to permit the use of proportionality constants in the functional relations. Assume that the variables of main concern are the armament levels in the two countries expressed in mega-dollars or mega-tons. These are time-dependent and will be denoted by $X_1(t)$ and $X_2(t)$ respectively. As usual, these levels will be the accumulation of their rates of change, so that the first task is to form a rate equation which describes the behavior of each level.

As in any modeling problem, the model can be no better than our understanding of the underlying situation and our knowledge of the validity of any approximations. For this particular example, we draw upon an earlier study to gain some insight.* The rate of change of each country's armament expenditures (and, with it, growth) depends largely on the perceived level of the nation and the view of its inhabitants on the raising and lowering of taxes which simply expresses the willingness to commit resources to an arms build-up. To implement these concepts the following model is hypothesized

$$\frac{\Delta X_1(t)}{\Delta t} = -k_{11}X_1(t) + k_{12}X_2(t) + U_1(t) \qquad (41)$$

and

$$\frac{\Delta X_2(t)}{\Delta t} = +k_{21}X_1(t) - k_{22}X_2(t) + U_2(t) \qquad (42)$$

where k_{ij} is the positive proportionality constant relating the j-th level to the rate of change in the i-th level. U_i is an external input (in units of mega-tons per month) which directly effects the rate of change of the i-th level. For example, U_1 may be taken as the current militaristic stance of country one. If a pacifist government is in power, U_1 may be negative indicating unilateral disarmament. According to Eq. (41), as $X_1(t)$ tends to grow**, the rate of change of level X_1 decreases, indicating a resistance to further taxation or an unwillingness to forego consumer goods for guns. Here both the influence and the level relate to the same country, so that the

---

* H. F. Richardson, Arms and Insecurity, Stevens & Sons, 1960

** Recall that in any linearization the variables are defined in terms of small changes about a nominal operating point.

proportionality constant k_{ij} becomes k_{11} for $X_1(t)$. On the other hand, as the armaments level in the other country, $X_2(t)$, grows, national defense policies may dictate an increase in the armaments level $X_1(t)$ of the first country. Here the subscript of the proportionality constant is k_{12}, i.e. the level in country one is affected by the level of country two. By symmetry, the same equation applies to the other country, although, of course, k_{12} is generally not equal to k_{21}.

The difference between Eq. (41) and the first order equations discussed previously, is the appearance of a new time-dependent variable, $X_2(t)$, in the $X_1(t)$ equation. Obviously, something must be known about $X_2(t)$ before Eq. (41) can be solved. The inverse applies to Eq. (42). Accordingly, the two equations are coupled and must be solved together. Mathematically they are known as a set of simultaneous, coupled, forst-order, difference equations.

If the time interval, Δt, in the two equations is allowed to become very small (and, in the limit, go to zero), then $\frac{\Delta}{\Delta t}$ can be replaced by the differential operator $\frac{d}{dt}$ and a set of simultaneous, coupled, first-order, differential equations will result:

$$\frac{dX_1(t)}{dt} = -k_{11}X_1(t) + k_{12}X_2(t) + U_1(t) \tag{43}$$

and

$$\frac{dX_2(t)}{dt} = +k_{21}X_1(t) - k_{22}X_2(t) + U_2(t) \tag{44}$$

The problem would become only slightly more complicated if a third country is added to our considerations of the arms race. Here a term $k_{13}X_3(t)$ would have to be added in the first equation, the second equation would also increase by one term and a new third equation would have to be included, leading to three simultaneous first order equations.

A different form of themathematical model for this problem can be derived from the results already obtained. First differentiate one equation, say Eq. (43), term by term with respect to time* to obtain:

$$\frac{d^2X_1(t)}{dt^2} = -k_{11}\cdot\frac{dX_1(t)}{dt} + k_{12}\cdot\frac{dX_2(t)}{dt} + \frac{dU_1(t)}{dt} \tag{45}$$

Now, substitute the other equation (Eq. (44)) into Eq. (45) to eliminate $dX_2(t)/dt$:

$$\frac{d^2X_1(t)}{dt^2} = -k_{11}\cdot\frac{dX_1(t)}{dt} + k_{12}[+k_{21}X_1(t) - k_{22}X_2(t) + U_2(t)] + \frac{dU_1(t)}{dt} \tag{46}$$

We next solve Eq. (43) for $X_2(t)$,

$$X_2(t) = \frac{1}{k_{12}}\cdot\frac{dX_1(t)}{dt} + \frac{1}{k_{12}}\cdot k_{11}X_1(t) - \frac{1}{k_{12}}\cdot U_1(t) \tag{47}$$

and substitute the result into Eq. (46) to eliminate $X_2(t)$. The resulting equation (after some algebraic rearrangement),

$$\frac{d^2X_1(t)}{dt^2} + (k_{11}+k_{22})\cdot\frac{dX_1(t)}{dt} + (k_{11}k_{22} - k_{12}k_{21})X_1(t) = k_{12}U_2(t) - k_{22}U_1(t) + \frac{dU_1(t)}{dt}, \tag{48}$$

now is a single, second order differential equation in $X_1(t)$ only and it no longer appears to be coupled to $X_2(t)$. If the operations are interchanged, i.e. Eq. (44) is first differentiated, Eq. (45) is substituted and Eq. (44) is used to eliminate $X_1(t)$, a corresponding equation can be found for $X_2(t)$

$$\frac{d^2X_2(t)}{dt^2} + (k_{11}+k_{22})\cdot\frac{dX_2(t)}{dt} + (k_{11}k_{22} - k_{12}k_{21})X_2(t) = -k_{21}U_1(t) + k_{11}U_2(t) + \frac{dU_2(t)}{dt}. \tag{49}$$

---

* From previous discussion, it is clear that differentiating $X_1(t)$ with respect to time yields dX_1/dt, which represents the rate of change of X_1. By direct extension, another application of the operator d/dt yields the "rate of change of the rate of change". The resulting expression d/dt ($dX_1(t)/dt$) is conventionally written as $d^2X_1(t)/dt^2$ and read as "d-squared X_1 of t - d-t squared" or, more simply as the second derivative of $X_1(t)$ with regard to t. An example of a second derivative is the acceleration which is the time rate of

The important thing to note is that two coupled first order equations have evolved into a single second order differential equation (either Eq. (48) or Eq. (49)) and that both forms contain the same information and lead to the same force-free solution.

Writing both Eq's. (48) and (49) in operator form where D^2 denotes the second derivative, their left hand homogenous sides can be expressed as

$$D^2 + (k_{11} + k_{22})\,D + (k_{11}\,k_{22} - k_{12}\,k_{21}) = 0 \qquad (50)$$

where the above expression operates on either $X_1(t)$ or $X_2(t)$. This is known as the characteristic equation, a point to which we will return in the discussion of the solution of second order systems.

It may be noted that, in general, n-coupled first order differential equations are equivalent to one n-th order uncoupled differential equation and both forms will supply us with the same information about the dynamic response of the system. While we have here only shown how to convert from the n coupled first order differential equations to a single n-th order differential equation, there are also available techniques (beyond the scope of these notes) for converting in the opposite direction -- from a single n-th order differential equation to n coupled first order differential equations. Incidentally, when an n-th order system is described by n coupled first order differential equations, the resulting n dependent variables are often called the state variables of the system, in that knowledge about those variables at any instant of time should completely define the future state of the system.

As another example of second order system modelling, consider the dynamics of the propagation of an epidemic* in an isolated city of total population N. Let I(t) be the number of persons infected at any instant, S(t) be the number of persons susceptible to infection at any instant, and R(t) be the instantaneous level of people who are recovered (or dead), and therefore now immune. Since the sum of the components must equal the whole, we have

$$N = S(t) + I(t) + R(t) \qquad (51)$$

The problem to be investigated is what happens in this city due to an initial level of I_0 people infected at time zero.

Medical evidence suggests the rate of change of the number of infected people is proportional to the product of the instantaneous levels of susceptible and infected people minus the rate of change of people recovering. This multiplicative relation will lead to a nonlinear mathematical model, which can be expressed as

$$\frac{dI(t)}{dt} = k_1 S(t)\,I(t) - \frac{dR(t)}{dt} \qquad (52)$$

Solving Eq. (51) for S(t) and substituting the result into Eq. (52) yields:

$$\frac{dI(t)}{dt} = k_1[\,N\,I(t) - I^2(t) - R(t)\,I(t)\,] - \frac{dR(t)}{dt} \qquad (53)$$

Past experience has also led to the observation that the rate of change of the level of recovered people is proportional to the instantaneous level of people infected (i.e., a certain fraction of sick people will get better every day). Mathematically, this is stated as

$$\frac{dR(t)}{dt} = k_2 I(t) \qquad (54)$$

Equations (53) and (54) comprise the two coupled first order differential equations which describe the epidemic dynamics.

---

change of velocity which in turn is the derivative of distance with regard to time. Acceleration, is therefore, the second derivative of distance with regard to time. (FOOTNOTE CONCLUDED FROM PREVIOUS PAGE)

* For a detailed discussion of this problem and a somewhat different model, see "The Man-Made World", Engineering Concepts Curriculum Program. McGraw-Hill, 1971. Pages 354 to 365 include a numerical solution.

The two primary state variables are the instantaneous levels of infected and recovered people.

Equation (53) is obviously nonlinear, and therefore requires special treatment. It involves the square of one of the dependent variables and the product of the two dependent variables. While it is mathematically possible to linearize Eq. (53), by an extension of the methods discussed at the end of Chapter 2, the question remains whether the linearized equations are applicable to the real problem. They will only be valid for small deviations about the nominal levels at which the linearization is performed. Accordingly, the results of solving such a linearized set of differential equations would only be good for a short period of time, perhaps a day or two of the epidemic dynamics, and then a new linearized model would have to be developed for further valid predictions. Since there are few analytical techniques for non-linear equations, alternate methods of solution must be found such as a numerical approach or an analog simulation. See the cited reference on this problem for further discussion and the next section of this chapter for an analog simulation of the problem.

If the sickness in this epidemic happens to have the characteristic that everyone infected recovers in exactly the same elapsed time, T, then an important simplification in the model may be made. Equation (54) may be replaced with

$$R(t) = I(t - T) \tag{55}$$

which says that the instantaneous level of recovered people is equal to the number of people infected T days ago. This is not a rate equation, and need not be integrated for a solution. It is termed a delayed function. Substituting Eq. (55) into Eq. (53), leads to

$$\frac{dI(t)}{dt} = k_1 [N \cdot I(t) - I^2(t) - I(t - T) \cdot I(t)] - k_2 I(t - T) \tag{56}$$

a single first order (nonlinear) differential equation which now describes the problem instead of the two coupled first order differential equations (52) and (53). In general, when modeling a dynamic system, each rate equation which may be replaced by a delayed function will usually reduce the order of the system model by one. More precisely, each delayed function introduced as a substitute for a rate equation will reduce by one the number of integrations necessary to solve the problem. Here, the concept of system order and the number of integrations needed to solve the system are used interchangeably.

In Eq. (56) the last term in the brackets is just the product of those infected today and those infected T days ago (for the first T days of the solution this term will be zero). Equation (56) is virtually intractable to an analytical solution, but is reasonably straightforward if attacked by numerical techniques as by analog simulation. For an analog simulation a unit not yet defined - the multiplier - will be needed as discussed in the next section. The lack of ability to solve the problem analytically is a normal state of affairs for most real systems, although we should repeat that often linearized mathematical models can offer us important insights even if not a complete solution.

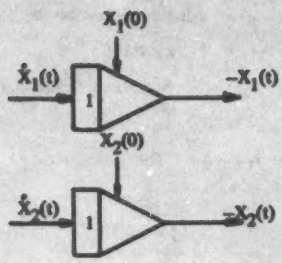
Analog Simulation of Second Order Systems

To develop the analog simulation of two coupled first order differential equations, it is advisable to first write the equations in the form of Eqs. (43) and (44) repeated here for convenience

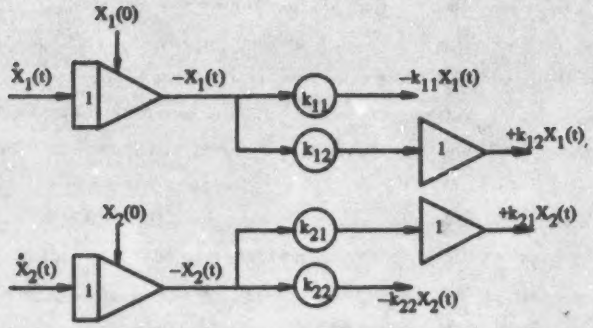
$$\frac{dX_1(t)}{dt} = -k_{11} X_1(t) + k_{12}(t) + U_1(t) \tag{43}$$

$$\frac{dX_2(t)}{dt} = +k_{21} X_1(t) - k_{22} X_2(t) + U_2(t) \tag{44}$$

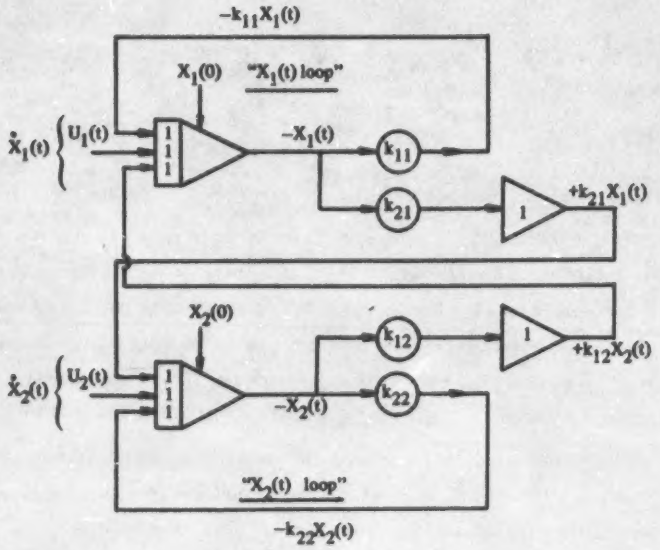
where the derivatives are isolated on the left hand side. Just as in the case of a first order system analog simulation, the procedure is to start by assuming that the derivatives are available so that operational amplifiers may be used in "boot-strap" fashion. Thus, in Fig. 21-(a), the rates of change of the two armaments levels, $\dot{X}_1(t)$ and $\dot{X}_2(t)$, are assumed to exist and are fed into separate integrating op-amps. Here and in the figure the short hand notation $\dot{X}(t)$ is used for $dX(t)/dt$. Also fed into the two integrators



(a)



(b)



(c)

Fig. 21 Analog simulation of two coupled first order differential equations.

are the initial conditions, $X_1(0)$ and $X_2(0)$. Since two differential equations must be solved, two integrators are required. In general, a separate integrator will be needed for each first order differential equation of the coupled set. The outputs of the integrators will be $-X_1(t)$ and $-X_2(t)$ as shown in the figure.

Referring to Eqs. (43) and (44), the rates are seen to include various constants multiplying the levels. Since the levels are now available (see Fig. 21-(a)), the next step is merely the multiplication of the levels by the appropriate constants as indicated in Fig. 21-(b). According to Eqs. (43) and (44), plus $k_{12}X_2(t)$ and plus $k_{21}X_1(t)$ are required to reconstitute the desired rates on the left hand side of the equations. However, these levels are available so far only with a minus sign in front of them (due to the sign inversion properties of the op-amp), so an additional sign inversion is needed. This is accomplished with the two inverting op-amps of unity gain as shown in Fig. 21-(b).

The four voltage signals now available at the right hand end of Fig. 21-(b) simulate four of the six terms on the right side of Eqs. (43) and (44). They are looped back and plugged into the input summing side of the integrating op-amps as shown in Fig. 21-(c). The remaining two signals, $U_1(t)$ and $U_2(t)$, the external inputs to the system, are supplied from an external signal voltage generator and may be arbitrarily prescribed functions of time. For example, on an analog computer they can readily be fed in as positive or negative step voltage changes, sinusoidally oscillating signals, zero, etc. depending on the assumed real situation. Fig. 21-(c) thus constitutes the complete analog simulation of Eqs. (43) and (44).

The voltages at various points in the patching diagram of Fig. 21-(c) are analogous to the modelled variables labelled at those points. As before, the precise correspondence depends upon the scaling factors chosen. If the maximum expected level $X_1(t)$ of armaments in country one is about 50 megatons, and the analog computer used is a ten volt machine, then a reasonable scaling factor would be to let one volt represent 5 megatons. A measured signal of -8.5 volts at the output of the top integrator in Fig. 21-(c) would thus be analogous to +42.5 megatons of armaments in country one at the instant of measurement. Each second of machine time may be conveniently viewed as, say, one month of real time to speed up the solution.

When the problem has been formulated as a single second order differential equation, as in the case of Eqs. (48) or (49), the analog simulation procedure is only slightly modified. Again, the starting point is to isolate the highest derivative on the left hand side of the equation so that Eq. (48), for example, would be rewritten as

$$\frac{d^2X_1(t)}{dt^2} = -(k_{11}+k_{22})\frac{dX_1(t)}{dt} - (k_{11}k_{22}-k_{12}k_{21})X_1(t) + k_{12}U_2(t) - k_{22}U_1(t) + \frac{dU_1(t)}{dt} \qquad (57)$$

Assuming the availability of $d^2X_1(t)/dt^2$, written as $\ddot{X}_1(t)$ for convenience, a voltage signal proportional to it is fed into an integrating op-amp as in Fig. 22-(a). The output of the integrator is $-d\dot{X}_1(t)/dt$ or $-\dot{X}_1(t)$. In Fig. 22-(b) two separate operations are indicated to formulate the right hand side of Eq. (57): first the signal $-\dot{X}_1(t)$ is multiplied by $(k_{11} + k_{22})$ and fed backwards, and secondly the same signal is fed forwards through an integrating op-amp to product the signal $+X_1(t)$ at its output. The next step is to multiply the $+X_1(t)$ signal by $(k_{11}k_{22} - k_{12}k_{21})$ and then pass it through a sign inverting op-amp of unity gain. As shown at the bottom of Fig. 22-(c), this generates a voltage which is proportional to the second term on the right hand side of Eq. (57).

To complete the analog simulation of Eq. (57) the external voltages proportional to the last three terms in the equation: $k_{12}U_2(t)$, $-k_{22}U_1(t)$, and $dU_1(t)/dt$ must be supplied. These are all arbitrary functions of time, and their voltage magnitudes depend upon the scaling factors chosen and the magnitudes of the actual physical values. Fig. 22-(d) shows the complete analog simulation of Eq. (48). A similar analog simulation may be developed for the second order differential equation in $X_2(t)$, Eq. (49).

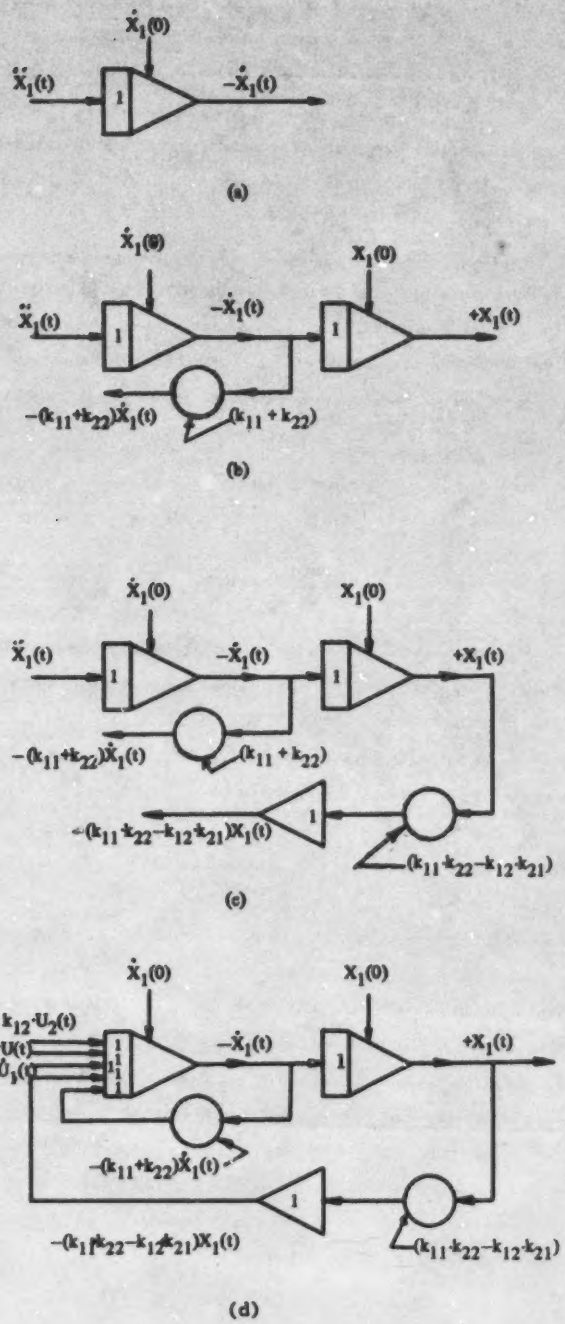


Fig. 22 Analog simulation of a second order differential equation.

Comparing the analog simulation of the two coupled first order differential equations (43) and (44), shown in Fig. 21-(c), to the analog simulation of the equivalent second order differential equation (48), shown in Fig. 22-(c), it should be noted that in either case two integrating op-amps are required. The coupled nature of the two simultaneous first order differential equations (43) and (44) is obvious by inspection of Fig. 21-(c) where the patching of the two integrating loops is physically coupled by jumper wires. On the other hand, Fig. 22-(d) contains information only about $X_1(t)$, which is to be expected since $X_2(t)$ was purposely eliminated from Eq. (48) which is simulated in the figure. While this simulation does not contain the information about $X_2(t)$ that Fig. 21-(c) does, it does explicitly provide the first derivative of $X_1(t)$ and one additional piece of information which is sometimes useful: the second derivative with respect to time of $X_1(t)$. (The voltage proportional to $d^2X_1(t)/dt^2$ occurs at the input to the left hand integrator in Fig. 22-(d). To explicitly obtain that voltage, two unity–gain inversion op-amps would have to be inserted before that integrator; the first would act as a summer and the output of the second would be the desired voltage.) In general, however, the form of the simulation to follow is usually dictated by the form of the mathematical model derived.

As a final example of analog simulation, consider the epidemic model discussed at the end of the previous section. The two coupled nonlinear first order differential equations (53) and (54), which were derived earlier to model the situation , are repeated here for convenience:

$$\frac{dI(t)}{dt} = k_1 \cdot [\, N \cdot I(t) - I^2(t) - R(t) \cdot I(t) \,] - k_2 I(t) \qquad (53)$$

and

$$\frac{dR(t)}{dt} = k_2 I(t). \qquad (54)$$

Figures 23-(a) and (b) are similar to Figures 21-(a) and (b), and the rationale behind them is much the same. However, a new element is introduced in Fig. 23-(c), the multiplier – represented by a small square box with a cross in it on the figure. The need for two multipliers should be evident from the last two terms in the brackets of Eq. (53). The multiplier takes as its input any two time varying voltages, and produces an output which is the negative scaled product of the two. The scaling consists of dividing by a constant, either ten or one hundred, depending on whether the computer is a ten or one hundred volt machine, to keep the output voltage within the limits of the machine.

The final simulation of the epidemic dynamics model is shown in Fig. 23-(d), where the four fed back inputs to the top integrator in the figure are analogous to the four terms on the right hand side of Eq. (53). Once again, the problem has been simulated by assuming the existence of the highest derivatives, and then proceeding to construct them. It makes little difference that the problem is represented by a nonlinear model except that the scaling is somewhat more complex and must be handled with caution. In all three of the examples considered in this section, the problem start is determined by the initial conditions supplied to the integrators. As soon as the "start" button is depressed (defined as time t = 0), the levels in the problem are set equal to the supplied initial values. As in the case of the first order system analog simulations, ensuing values of the levels may be either read directly with a voltmeter by depressing a "hold" button at the instant the reading is desired, or the levels may be continuously recorded on a multi-channel recording device. Such a recording is presented in Figure 24 for N = 52,000, I(0) = 1,000 and R(0) = 0. These numbers in the previously cited reference on this problem concur with the numerical values, and the result shown in Fig. 24 coincides with the plot of the numerical solution reported there. Note how the number of infected, I(t), first grows and then decays with time, never reaching N as in the real situation.

Analysis of Second Order Systems

Earlier in this chapter, a second order characteristic equation (50) was arrived at which represented the dynamics of the armaments level problem for both countries one and two. In simplest form this characteristic equation may be rewritten as

$$D^2 + b \cdot D + c = 0 \qquad (58)$$

where for Eq. (50), $b = (k_{11} + k_{22})$, and $c = (k_{11} \cdot k_{22} - k_{12} \cdot k_{21})$. Note that no generality is lost by taking the

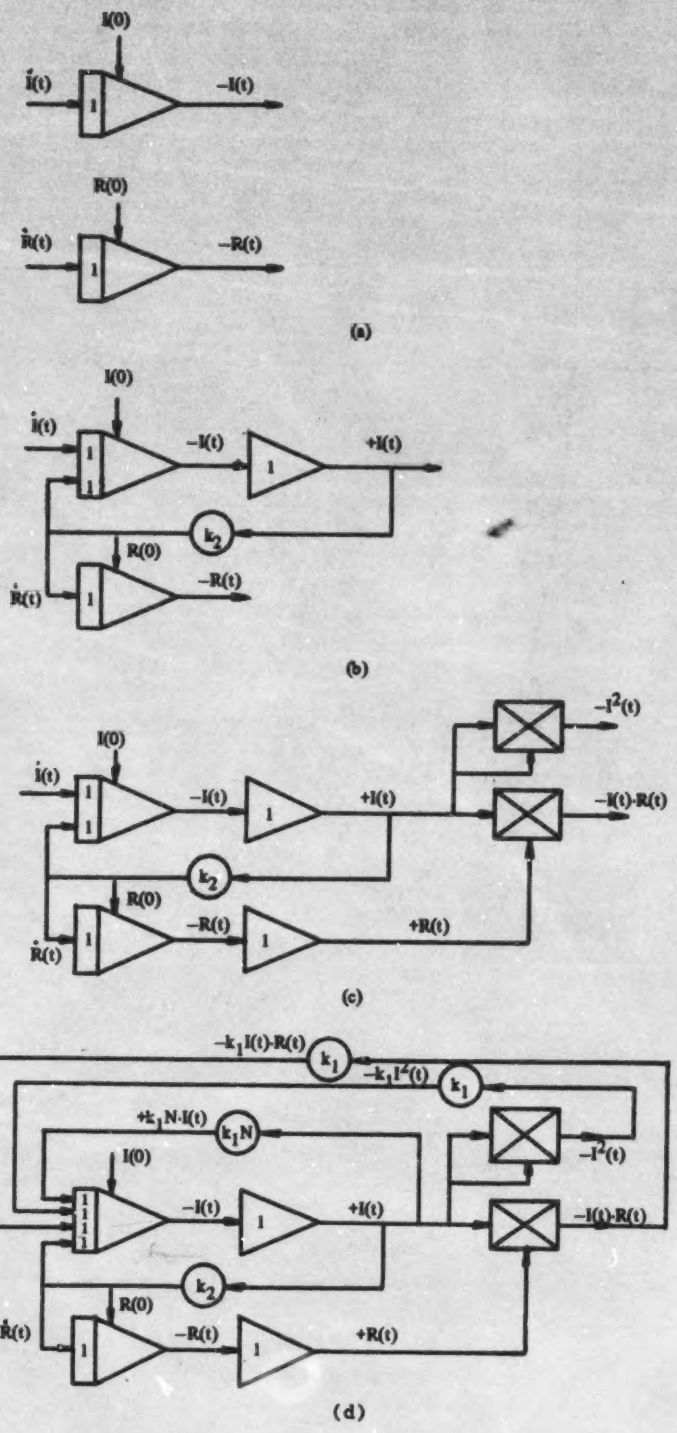


Fig. 23 Analog simulation of a second order non-linear model for epidemic dynamics.

47

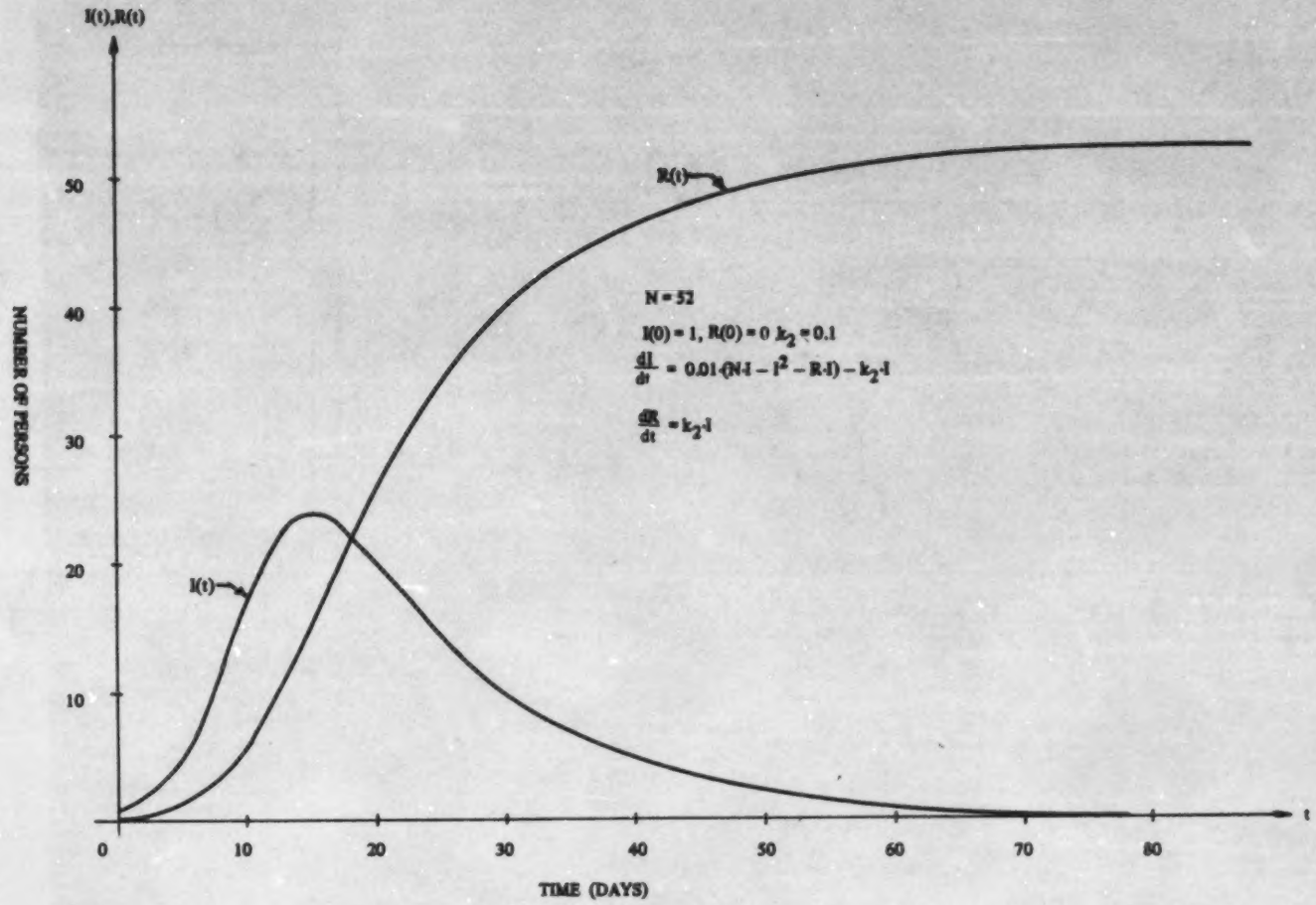


Fig. 24 Analog computer output of the epidemic dynamics problem.

coefficient of the D^2 term in Eq. (58) as unity, since a non-unity value could be divided through, bringing us back to the form of Eq. (58) with newly defined constants b and c. Recall that the characteristic equation is derived from the original system differential equation by:

a) replacing the derivative operation by a notational operator, D,

b) deleting the variable, $X_1(t)$ or $X_2(t)$ in this case,

and c) setting the result equal to zero.

As will be shown below, the characteristic equation supplies a good deal of information about the dynamics of the system it describes. The nature of the similarities between the two responses is well described by the system's characteristic equation, and it is this nature that we next investigate. However, we must keep in mind that the different nonzero right hand sides of Eqs. (48) and (49) also lead to the differences in the responses of $X_1(t)$ and $X_2(t)$.

Treating Eq. (58) like an algebraic equation, it is a simple quadratic in D and the two roots are

$$D_{1,2} = -\frac{b}{2} \pm \frac{1}{2}\sqrt{(b^2 - 4c)} \tag{59}$$

where D_1 is associated with the plus sign, and D_2 with the minus sign. The roots of the characteristic equation, usually called the system's eigenvalues, determine the nature of the system's transient response. . . just as in the first order case of the previous chapter.

Several possibilities arise: if in Eq. (59) $b^2 > (4c)$, then the two roots $D_{1,2}$ will be real and unequal. If $b^2 = (4c)$, then the roots will be real and both equal to $-(b/2)$. And finally, if $b^2 < (4c)$, then the roots will be complex conjugate*, given by

$$D_{1,2} = -\frac{b}{2} \pm \frac{j}{2}\sqrt{(4c - b^2)} \tag{60}$$

The second case of real and precisely equal roots seldom occurs in physical problems and is mainly of academic interest. However, the first and third cases have physical significance and we will consider them in turn.

When $b^2 > (4c)$ and the roots D_1 and D_2 are real and unequal, the solution is relatively straightforward. The same reasoning that led to Eq. (32) as the analytical form of the first order transient solution leads to a similar proposed solution here, namely

$$X(t) = C_1 e^{D_1 t} + C_2 e^{D_2 t} \tag{61}$$

where e is the base of the natural logarithm and C_1 and C_2 are constants of integration which must be evaluated by using the information known from the initial conditions. The derivation of Eq. (61) is beyond the scope of these notes, but it may be shown by back substitution that Eq. (61) does indeed satisfy the homogeneous form of the original differential equation (48) or (49). We will here concentrate only on the information contained in the solution format of Eq. (61).

To evaluate the constants of integration C_1 and C_2, it is first necessary to write an expression for the rate of change of X(t), namely

$$\frac{dX(t)}{dt} = D_1 C_1 e^{D_1 t} + D_2 C_2 e^{D_2 t} \tag{62}$$

It is now possible to evaluate C_1 and C_2 as long as X(0) and $dX(0)/dt$ are both known, i.e., as long as the initial value of X(t) and the initial value of $dX(t)/dt$ are given. The problem now reduces to solving two simultaneous algebraic equations for two unknowns. An example should clarify the procedure.

---

* Complex conjugate numbers have equal real parts and imaginary parts that differ only in sign. Thus $(-3 + 4j)$ and $(-3 - 4j)$ are complex conjugate, where $j = \sqrt{-1}$

In Eq. (50), let $k_{11} = 2$, $k_{22} = 4$, $k_{12} = 1$, and $k_{21} = 3$; so that $b = 6$, and $c = 5$ in Eq. (58). Accordingly, the eigenvalues become, from Eq. (59)

$$D_{1,2} = -1, -5. \tag{63}$$

The free and transient solution (i.e., without driving or forcing functions on the right hand side of the equations) for either $X_1(t)$ or $X_2(t)$ was given by Eq. (61) as

$$X(t) = C_1 e^{-t} + C_2 e^{-5t}.$$

To go further we must focus on either $X_1(t)$ or $X_2(t)$. If it is assumed that $X_1(0) = +2$ and $dX_1(0)/dt = 0$, substituting these initial conditions into Eqs. (61) and (62) and evaluating conditions at $t = 0$ yields:

$$2 = C_1 + C_2$$

and

$$0 = -C_1 - 5C_2.$$

Solving these two simultaneous algebraic equations leads to $C_1 = +2.5$ and $C_2 = -0.5$. Thus, the free and transient solution, with no forcing function, becomes:

$$X_1(t) = +2.5\,e^{-t} - 0.5\,e^{-5t}. \tag{64}$$

The transient solution for $X_2(t)$ is also immediately at hand. For the values already given, and taking $U_1(t) = U_2(t) = 0$, to get the force free response, Eqs. (43) and (44) become (at $t = 0$):

$$0 = -2\cdot2 + X_2(0)$$

and

$$\frac{dX_2(0)}{dt} = 3\cdot2 - 4X_2(0).$$

Solving, $X_2(0) = +4$ and $dX_2(0)/dt = -10$, and substituting these initial conditions into Eqs. (61) and (62) yields:

$$4 = C_1 + C_2$$

and

$$-10 = -C_1 - 5C_2$$

from which $C_2 = +1.5$ and $C_1 = +2.5$. The solution for $X_2(t)$ thus is:

$$X_2(t) = +2.5\,e^{-t} + 1.5\,e^{-5t}. \tag{65}$$

That the second order differential equation (48) in $X_1(t)$, after having been solved for the transient solution of Eq. (64), leads immediately – through some simple algebra – to a solution for $X_2(t)$ should not be surprising. It is simply a reaffirmation of the basic coupling between $X_1(t)$ and $X_2(t)$ and the fact that Eq. (48) is indeed equivalent to the pair of equations (43) and (44). Again, it is possible to reverse this whole procedure by starting with Eq. (49) and the above initial conditions on $X_2(t)$, and then arriving at the same results of Eqs. (64) and (65).

Before examining the second case of complex conjugate eigenvalues, let us interpret Eqs. (64) and (65) in some detail. They are plotted in Fig. 25 (in the next section). First, it is important to note that either solution contains two modes of response, an e^{-t} term and an e^{-5t} term, and that these two modes (the eigenmodes) appear in the solutions for both $X_1(t)$ and $X_2(t)$. These eigenmodes derive directly from the eigenvalues which are determined from the system characteristic equation. The amount of each eigenmode, as determined by the premultiplying constant, appearing in each solution, however, will be different (compare Eqs. (64) and (65)!), and will depend upon the initial values. Second, we note that the solution of a second order system will require two integrations and, therefore, two constants of integration. Accordingly two initial conditions must be supplied to evaluate those constants. Thirdly, real roots of the characteristic equation again lead to an exponential form of the solution, similar to that observed in the previous chapter, except that now the solution is made up of the sum of two exponentials. In general, with two

different roots, the response will consist of a portion (or eigenmode) that proceeds slowly, here e^{-t}, and one that proceeds fast, here e^{-5t} which decays much more rapidly than e^{-t}. Finally, observe that had the eigenvalues been positive, the response would have indicated an exponential growth rather than an exponential decay. For our model of the armaments race with certain numerical values of k_{ij}, the analysis could predict a run-away spiralling arms build-up.

At this point in the previous chapter on first order systems, the response to a forcing function (i.e., non-zero terms on the right hand side of the system differential equation) was considered. Such a tack at this point, while interesting and informative, would not add significantly to our understanding of general transient system behavior. Suffice it to say that forcing functions, as before, lead to the so-called particular or steady-state solution of the differential equation which usually results in a new system equilibrium point. The procedures for finding the second order forced response are not too different from those discussed in the previous chapter. But forcing functions do not affect the nature of the system's transient response. This is found from the system characteristic equation and determines whether the system will exponentially grow or decay, and how fast it will respond. The forced response will merely follow the forcing function in both time and form.

Let us now consider the second case of interest where the eigenvalues are complex conjugate, as in Eq. (60). In order to discuss this more succinctly, choose at the outset numerical values of $k_{11} = +2$, $k_{22} = +2$, $k_{12} = +4$ and $k_{21} = -4$ (this last negative value indicates leaders in country two have for some reason decided to decrease their armaments, while they perceive the level of armaments in country one increasing!). For these k_{ij} values, $b = +4$ and $c = +20$. Eq. (60) becomes

$$D_{1,2} = -2 \pm 4j \tag{66}$$

The general solution of Eq. (61) becomes:

$$X(t) = C_1 e^{(-2+4j)t} + C_2 e^{(-2-4j)t}$$

or,

$$X(t) = e^{-2t} [\, C_1 e^{4jt} + C_2 e^{-4jt} \,] . \tag{67}$$

The factored-out exponential term e^{-2t} shows that whatever the bracketted term becomes, it will be multiplied by a decaying exponential. This will always be true as long as the real part of the eigenvalue (-2 in Eq. (66)) is negative. The reverse is also true – a positive real part of the complex conjugate eigenvalues will lead to a solution containing a multiplicative term which grows exponentially.

To understand the nature of the bracketted term on the right of Eq. (67), first consider the Euler equation:

$$e^{\pm j\theta} = \cos\theta \pm j \sin\theta \tag{68}$$

which may be shown to be true by expanding each term in a Taylor series about $\theta = 0$. Taking θ as (4t) and using Eq. (68) in (67):

$$X(t) = e^{-2t} [\, C_1 (\cos 4t + j \sin 4t) + C_2 (\cos 4t - j \sin 4t)] .$$

Regrouping terms,

$$X(t) = e^{-2t} [\, (C_1 + C_2) \cos 4t + (C_1 j - C_2 j) \sin 4t]$$

or

$$X(t) = e^{-2t} [C_3 \cos 4t + C_4 \sin 4t] \tag{69}$$

where

$$C_3 = (C_1 + C_2)$$

and $$C_4 = (C_1 j - C_2 j)$$

are newly defined constants of integration*.

---

* Since X(t) is a real response, C_3 and C_4 must be real coefficients in Eq. (69). Thus, the previously defined constants C_1 and C_2 must be complex conjugate. In any event, these constants of integration have not yet been solved for, so replacing one set with another provides no loss in generality.

The transient response of Eq. (69) is seen to be a harmonic oscillation multiplied by an exponential decay. This response will die out with time. It is thus seen that the gamble of country two has paid off for this particular set of parameters. They have lowered their armaments expenditure and yet the levels of both $X_1(t)$ and $X_2(t)$ have levelled off after some damped oscillations. Since the oscillations may be quite large, as shown on Fig. 26, the original decision may have required considerable political courage! Fig. 26 appears in the next section, where both the analytical and numerical solutions for this problem are plotted together for comparison. Let us take the same initial conditions as in the case previously calculated for real eigenvalues and ascertain that this is indeed true.

To evaluate the constants of integration $dX(t)/dt$ is required, which is found to be:

$$\frac{dX(t)}{dt} = -2e^{-2t}[C_3\cos 4t + C_4\sin 4t] + e^{-2t}[-4C_3\sin 4t + 4C_4\cos 4t]. \quad (70)$$

Substituting $X_1(0) = +2$ and $dX_1(0)/dt = 0$ (as before) into Eqs. (69) and (70) evaluated at $t = 0$, we get:

$$2 = C_3$$

and

$$0 = -2C_3 + 4C_4$$

from which $C_3 = +2$ and $C_4 = +1$ are the constants of integration. The solution for $X_1(t)$ is then,

$$X_1(t) = e^{-2t}[2\cos 4t + \sin 4t] \quad (71)$$

Just as before, the transient solution for $X_2(t)$ may immediately be found by algebraic manipulation as

$$X_2(t) = e^{-2t}[\cos 4t - 2\sin 4t] \quad (72)$$

where $X_2(0) = +1$ and $dX_2(0)/dt = -10$ from Eqs. (43) and (44), and using $X_1(0) = +2$ and $dX_1(0)/dt = 0$. These last two equations comprise the analytical solution, and are plotted in Fig. 26 as already noted.

Eqs. (71) and (72) play the same role for the complex conjugate eigenvalue case as did Eqs. (64) and (65) for the case of real eigenvalues. The discussion in the paragraph following Eq. (65) again pertains except that the two separate eigenmodes are here identified as $(e^{-2t}\cos 4t)$ and $(e^{-2t}\sin 4t)$. Now there is, of course, no fast or slow mode of response – just a harmonically oscillating response which is damped out be a decaying exponential (as long as the eigenvalues have negative real parts).

Before leaving the topic of second order systems analysis, it should be noted that for the linear equation system discussed in this illustrative problem it is also possible to proceed to an analytical solution directly from the coupled first order differential Eqs. (43) and (44) instead of working with the second order differential Eq. (48) or (49). Without belaboring this point, which involves more advanced techniques, the procedures utilize the vector and matrix manipulations of linear algebra. To illustrate this, Eqs. (43) and (44) can be recast into vector format as:

$$\frac{d}{dt}\begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} = \begin{bmatrix} -k_{11} & +k_{12} \\ +k_{21} & -k_{22} \end{bmatrix} \cdot \begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} + \begin{bmatrix} U_1(t) \\ U_2(t) \end{bmatrix} \quad (73)$$

The same system characteristic equation (50) is embedded in this format as:

$$\text{determinant} \left\{ \begin{bmatrix} D & 0 \\ 0 & D \end{bmatrix} - \begin{bmatrix} -k_{11} & +k_{12} \\ +k_{21} & -k_{22} \end{bmatrix} \right\} = 0$$

or $(D + k_{11})(D + k_{22}) - k_{12}k_{21} = D^2 + (k_{11} + k_{22})D + (k_{11}k_{22} - k_{12}k_{21}) = 0.$

Numerical Analysis of Second Order Systems

The starting point for the numerical solution of two coupled first order differential equations is the expression of the problem in terms of an equivalent set of coupled difference equations. For the differential equation problems analyzed in the previous section, the equivalent difference equations are (41) and (42). These two equations are rewritten below, but with a

slightly different notation. In the previous chapter, subscripts denoted time indexing while there they denote the state variables. For time, parenthetical indexing will be used. Thus $X_1(i)$ is the value of X_1 at the i-th instant, and $X_2(i-1)$ is taken to be the value of X_2 at the $(i-1)$-th instant. Considering ΔX to be equal to $[X(i) - X(i-1)]$ in Eqs. (41) and (42), evaluating all terms on the right hand side of the equations at the $(i-1)$-th instant, multiplying through by Δt and rearranging,

$$X_1(i) = \Delta t\,[-k_{11}X_1(i-1) + k_{12}X_2(i-1) + U_1(i-1)] + X_1(i-1) \qquad (74)$$

and

$$X_2(i) = \Delta t\,[+k_{21}X_1(i-1) - k_{22}X_2(i-1) + U_2(i-1)] + X_2(i-1)\,. \qquad (75)$$

These two equations represent the computing algorithm for numerically solving the same problems which were dealt with analytically in the previous section. However, before proceeding, several steps are first required:

a) $U_1(t)$ and $U_2(t)$ must be made available in discretized form as functions of time

b) Δt, the computing interval, must be properly chosen

c) $X_1(0)$ and $X_2(0)$, the initial conditions, must be known

and d) the problem parameters (constants) must be numerically known.

To consider these items in turn, the continuous-in-time forcing functions, $U_1(t)$ and $U_2(t)$, are replaced by their staircase approximations as discussed earlier. To simplify matters in the example discussed below, the forcing functions will be set equal to zero to find the force-free transient solution which can be compared directly to the analytical solutions found previously. In the choice of the computing interval, Δt, something should be known about the nature of the transient since in the previous chapter it was indicated that the computing interval should be no larger than one half the system time constant. However, as evident from Eq. (64), there are now two time constants involved -- one and one-fifth of a month (from the two $e^{-t/T}$ terms in Eq. (64) -- where T = 1 and 0.2). The correct procedure is to apply this rule to the smallest system time constant, 0.2 month here, so that the computing interval will be chosen as 0.1 month. In the absence of such guiding information as the solution of Eq. (64), an educated guess must be used for the computing interval to arrive at a first solution. Then the analysis should be repeated, with a new interval equal to one half of the previous one to ascertain whether the solution is sensitive to the interval size. For the present purpose, we will use the computing interval of one-tenth of a month, and compare the results to the exact solutions of Eqs. (64) and (65).

For items (c) and (d) above, the same numerical values as given just before and after Eq. (63) will be used to permit a direct comparison between approaches. Thus, $k_{11} = 2$, $k_{22} = 4$, $k_{12} = 1$, $k_{21} = 3$, $X_1(0) = +2$, and $X_2(0) = +4$, and Eqs. (74) and (75) become:

$$X_1(i) = +0.8\,X_1(i-1) + 0.1\,X_2(i-1) \qquad (76)$$

and

$$X_2(i) = +0.3X_1(i-1) + 0.6\,X_2(i-1). \qquad (77)$$

With these two difference equations and the given initial values, the numerical solution proceeds as shown in Table 4.

In the table, the time index in the first column represents an elapsed time of one-tenth of a month per row, so that, for example, $X_2(12) = .71$ in the last column of the twelfth row, is the value of X_2 after 1.2 months. The second and third columns in any row represent the values of X_1 and X_2 at the previous tenth of a month which are transfered from the last two columns from the previous row. However, in the first row there is no transfer and $X_1(i-1)$ and $X_2(i-1)$ are the prescribed initial values. This transfer simply implies that the value at the start of any interval is the terminal value of the previous interval. The fourth through seventh columns of the table are simply the four terms on the right hand side of Eqs. (76) and (77), which are then used computationally to produce $X_1(i)$ and $X_2(i)$ in columns (8) and (9).

One striking result in Table 4 is that $X_1(i)$ and $X_2(i)$ become equal from about i = 10 (one month) onward, at least within the roundoff error of two significant figures. This should not be surprising, since examination of Eqs. (64) and (65) indicates that the e^{-5t} term is negligible in both equations from about t = 1.0 onward, leaving nearly identical results for both $X_1(t)$ and $X_2(t)$ in the analytical solutions.

TABLE 4

Δt = 0.10 month

1	2	3	4	5	6	7	8	9
i	$X_1(i-1)$	$X_2(i-1)$	$0.8X_1(i-1)$	$0.1X_2(i-1)$	$+0.3X_1(i-1)$	$+0.6X_2(i-1)$	$X_1(i)$	$X_2(i)$
1	$X_1(0)$=+2	$X_2(0)$=+4	1.60	.40	.60	2.40	2.00	3.00
2	2.00	3.00	1.60	.30	.60	1.80	1.90	2.40
3	1.90	2.40	1.52	.24	.57	1.44	1.76	2.01
4	1.76	2.01	1.41	.20	.53	1.21	1.61	1.74
5	1.61	1.74	1.29	.17	.48	1.04	1.46	1.52
6	1.46	1.52	1.17	.15	.44	.91	1.32	1.35
7	1.32	1.35	1.06	.13	.40	.81	1.19	1.21
8	1.19	1.21	.95	.12	.36	.73	1.07	1.09
9	1.07	1.09	.86	.11	.32	.65	.97	.97
10	.97	.97	.78	.10	.29	.58	.88	.87
11	.88	.87	.70	.09	.26	.52	.79	.78
12	.79	.78	.63	.08	.24	.47	.71	.71
13	.71	.71	.57	.07	.21	.43	.64	.64
14	.64	.64	.51	.06	.19	.38	.57	.57
15	.57	.57	.46	.06	.17	.34	.52	.51
16	.52	.51	.42	.05	.16	.31	.47	.47
17	.47	.47	.38	.05	.14	.28	.43	.42
18	.43	.42	.34	.04	.13	.25	.38	.38
19	.38	.38	.30	.04	.11	.23	.34	.34
20	.34	.34	.27	.03	.10	.20	.30	.30
21	.30	.30	.24	.03	.09	.18	.27	.27
22	.27	.27	.22	.03	.08	.16	.25	.24
23	.25	.24	.20	.02	.07	.15	.22	.22
24	.22	.22	.18	.02	.07	.14	.20	.21

The results of Table 4 are also plotted in Fig. 25. For comparison, the analytical solutions for $X_1(t)$ and $X_2(t)$ from Eqs. (64) and (65) are added. The agreement between the analytical and the numerical approach is within 10% at all times, so that the choice of a 0.1 month computing interval may be acceptable. Unfortunately, it is not possible to say whether the numerical solution will always fall above or below the exact analytical solution. Here it is seen that the numerical solution for X_1 starts out slightly high and then ends up slightly low, while the numerical solution for X_2 is low throughout.

As a second, and last numerical example, reconsider the previous case which led to complex conjugate eigenvalues and a damped oscillatory response. The constants used in that example were $k_{11} = +2$, $k_{22} = +2$, $k_{12} = +4$ and $k_{21} = -4$. Taking these values, considering only the force-free solution with $U_1(t)$ and $U_2(t) = 0$, and taking $\Delta t = 0.05$ months (see next paragraph), Eqs. (74) and (75) become:

$$X_1(i) = +0.9\,X_1(i-1) + 0.2X_2(i-1) \tag{78}$$

and

$$X_2(i) = -0.2X_1(i-1) + 0.9\,X_2(i-1)\ . \tag{79}$$

Using the same initial conditions as those leading to the analytical solution of Eqs (71) and (72), namely $X_1(0) = +2$ and $X_2(0) = +1$, Eqs. (78) and (79) are numerically solved in Table 5.

As an initial guess at the computing interval Δt, the e^{-2t} exponential multiplier term in Eqs. (71) and (72) indicates an effective time constant of 0.5 months – so the earlier rule of thumb would dictate a computing interval no larger than 0.25 months. However, an overriding consideration is the behavior of the oscillatory portion of the response which predicts a period T_p of

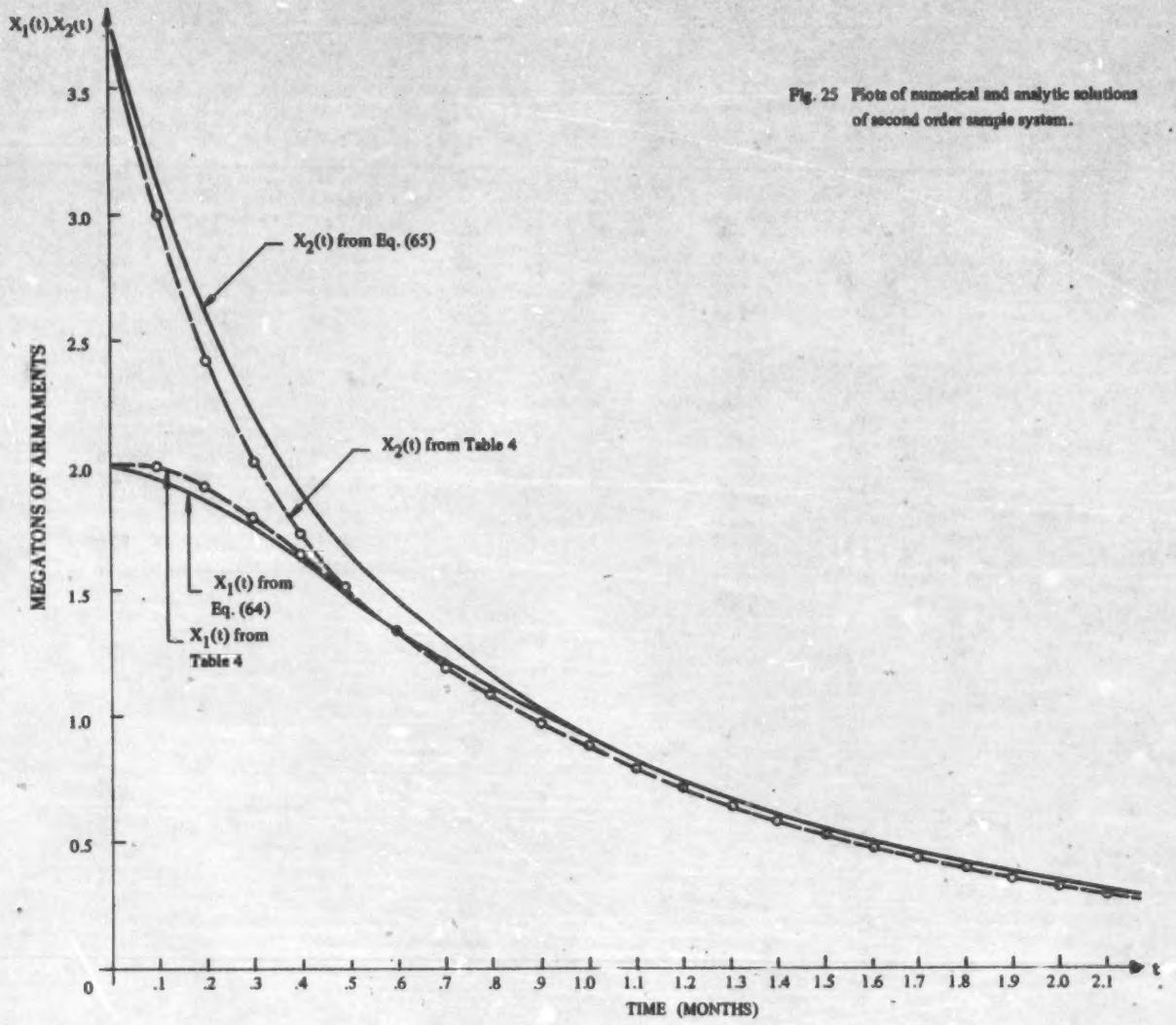


Fig. 25 Plots of numerical and analytic solutions of second order sample system.

$(\pi/2) = 1.57$ months (since one period will elapse when the argument in Eqs. (71) or (72) reaches $4T_p = 2\pi$). A computing interval of 0.25 months would therefore compute only six points per cycle – barely enough points to plot one cycle of a sinusoidal curve*.

As a first try at a numerical solution, a Δt of 0.20 months was attempted, based on the guidance provided by the analytical solution. In general, the choice of the correct Δt involves some trial and error as will be discussed below. The results for $\Delta t = 0.20$, not presented here, were totally unrealistic by leading to an undamped oscillatory solution. There is an inherent danger in using certain time intervals in a numerical approach which should predict a sinusoidal behavior and for that specific choice the (unstable) dynamics of the numerical solution technique overwhelmed the dynamics of the actual problem. To check against such an unusual occurrence, the numerical solution should be repeated with a Δt equal to one half of the previous run, and this procedure should be repeated until acceptable results are obtained. Accordingly, numerical solutions were recomputed with $\Delta t = 0.10$ month and then $\Delta t = 0.05$ month. This last run is tabulated in Table 5.

In this table, column (1) indicates the time index i, so each row represents 0.05 months. To save space, only 1.25 months (25 rows) are tabulated, while the first 2.5 months of the solution are plotted on Fig. 26. The format of Table 5 is similar to that

* In communications analysis there is a theorem due to Shannon which states that there must be at least two computations per cycle to have any meaningful results.

TABLE 5

Δt = 0.05 month

1	2	3	4	5	6	7	8	9
i	$X_1(i-1)$	$X_2(i-1)$	$0.9X_1(i-1)$	$0.2X_2(i-1)$	$-0.2X_1(i-1)$	$+0.9X_2(i-1)$	$X_1(i)$	$X_2(i)$
1	2.00	1.00	1.80	.20	−.40	.90	2.00	.50
2	2.00	.50	1.80	.10	−.40	.45	1.90	.05
3	1.90	.05	1.71	.01	−.38	.04	1.72	−.34
4	1.72	−.34	1.55	−.07	−.34	−.31	1.48	−.65
5	1.48	−.65	1.33	−.13	−.30	−.58	1.20	−.88
6	1.20	−.88	1.08	−.17	−.24	−.79	.91	−1.03
7	.91	−1.03	.82	−.21	−.18	−.93	.61	−1.11
8	.61	−1.11	.55	−.22	−.12	−1.00	.33	−1.12
9	.33	−1.12	.29	−.22	−.06	−1.01	.07	−1.07
10	.07	−1.07	.06	−.21	−.01	−.97	−.15	−.98
11	−.15	−.98	−.13	−.20	.03	−.88	−.33	−.85
12	−.33	−.85	−.30	−.17	.06	−.76	−.47	−.70
13	−.47	−.70	−.42	−.14	.09	−.63	−.56	−.54
14	−.56	−.54	−.50	−.11	.11	−.48	−.61	−.37
15	−.61	−.37	−.55	−.07	.12	−.33	−.62	−.21
16	−.62	−.21	−.56	−.04	.12	−.19	−.60	−.07
17	−.60	−.07	−.54	−.01	.12	−.06	−.55	.06
18	−.55	.06	−.50	.01	.11	.06	−.49	.17
19	−.49	.17	−.44	.03	.10	.15	−.41	.25
20	−.41	.25	−.37	.05	.08	.23	−.32	.31
21	−.32	.31	−.28	.06	.06	.28	−.22	.34
22	−.22	.34	−.20	.07	.04	.31	−.13	.35
23	−.13	.35	−.12	.07	.03	.31	−.05	.34
24	−.05	.34	−.05	.07	.01	.31	.02	.32
25	.02	.32	.02	.06	.00	.28	.08	.28

described for Table 4. The first row of columns (2) and (3) are the initial values of X_1 and X_2. Columns (8) and (9) are the values of $X_1(t)$ and $X_2(t)$ plotted in Fig. 26.

Figure 26 includes, for comparison purposes, both the analytical solutions of $X_1(t)$ and $X_2(t)$ from Eqs. (71) and (72) and the numerical solutions of $X_1(t)$ and $X_2(t)$ from Table 5. While a further decrease in the computing interval Δt would bring the numerical and the analytical solutions into closer correspondence, the present results may already be acceptable for most practical situations. As shown on the plot, the numerical solution predicts the correct periodicity and damping effects, and the quantitative values predicted are always within 10% of the analytical solution based on the initial values.

For a final check and comparison, Fig. 27 presents the problem solution as obtained by analog simulation. A computer simulation of the problem, following Fig. 21-(c), was run and the results recorded on a chart recorder. A trace of the actual recording is presented in Fig. 27, and the correspondence between Figs. 26 and 27 is evident.

Higher Order Systems

The dynamic response of higher than second order systems involves no new concepts – only additions to and complications of the previously discussed techniques. Here, several extensions and generalizations of the previous discussions are of interest.

The extension of analog simulation techniques to higher order systems should be the most obvious: additional op-amps will be required, and, in particular, the number of integrating op-amps must be equal to the system order. An initial condition must be supplied to each integrator. Scaling may become more complex for higher order systems, but the method described earlier still

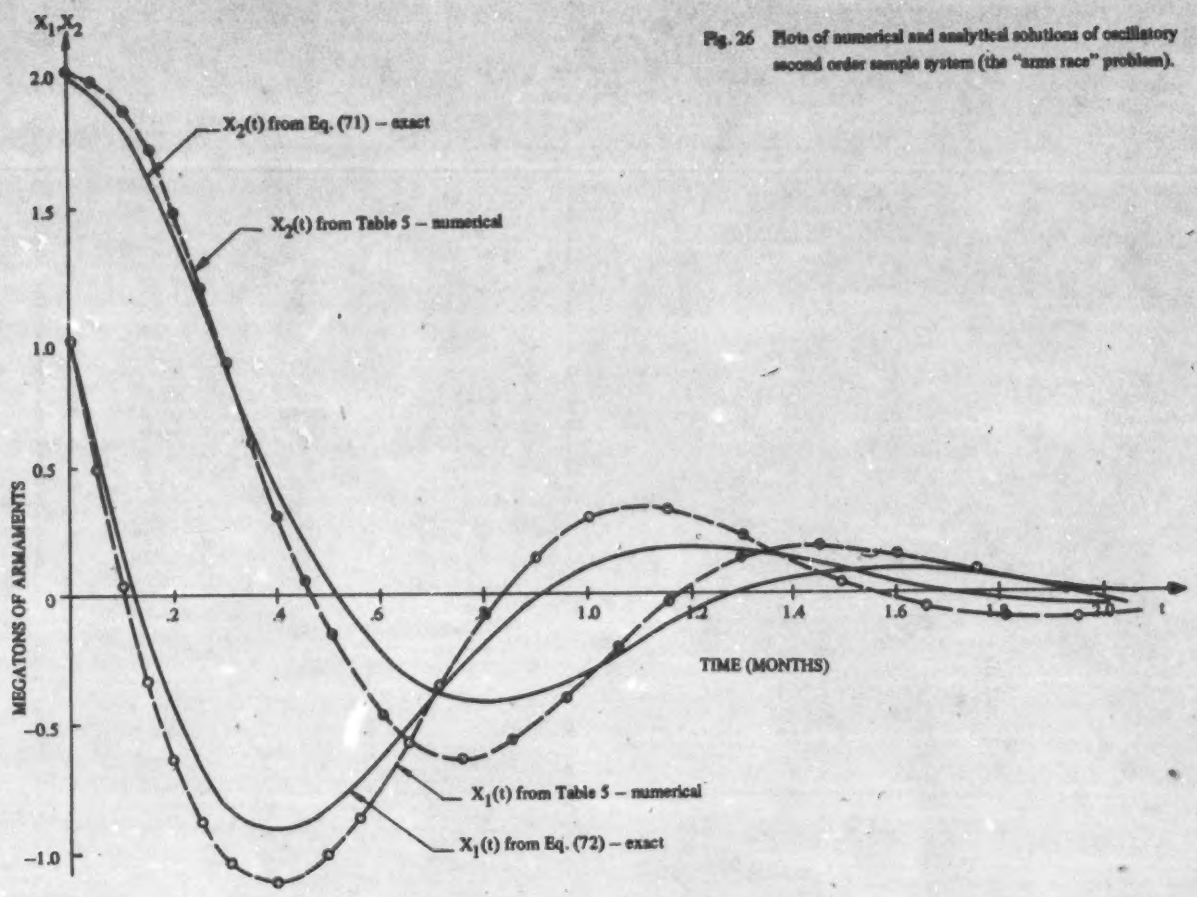


Fig. 26 Plots of numerical and analytical solutions of oscillatory second order sample system (the "arms race" problem).

pertains.

The second order system numerical solution techniques also extend easily to higher order systems. An n-th order system would be described by n coupled, first order difference equations, each one containing up to n terms. The numerical tabulation approach of Tables 4 and 5 extend directly, but the resulting tables will contain more columns. Again, without going into the mathematics in detail, the computing interval must be chosen small enough and should be checked by halving.

The mathematical analysis of higher order systems leads to perhaps the best understanding of what is involved. A second order system analysis led to the quadratic polynomial in the operator D given by Eq. (50). The roots of this quadratic were used in the two exponential terms of the solution given by Eq. (61). For higher order systems, analysis leads to n-th order polynomials in D having n roots, and n exponential terms in the solutions. Each real root of the system characteristic equation will lead to an exponential (decaying or growing) term, and each pair of complex conjugate roots will lead to a harmonic term which either decays or grows.

We thus see that the dynamic response of higher order systems does not add any fundamental change to the system behavior which is not already inherent in a second order system. The complications in the techniques arising from higher order systems are merely computational. Advanced studies in this area are essentially concerned with the development of rapid and accurate methodoligies for handling these computational complications.

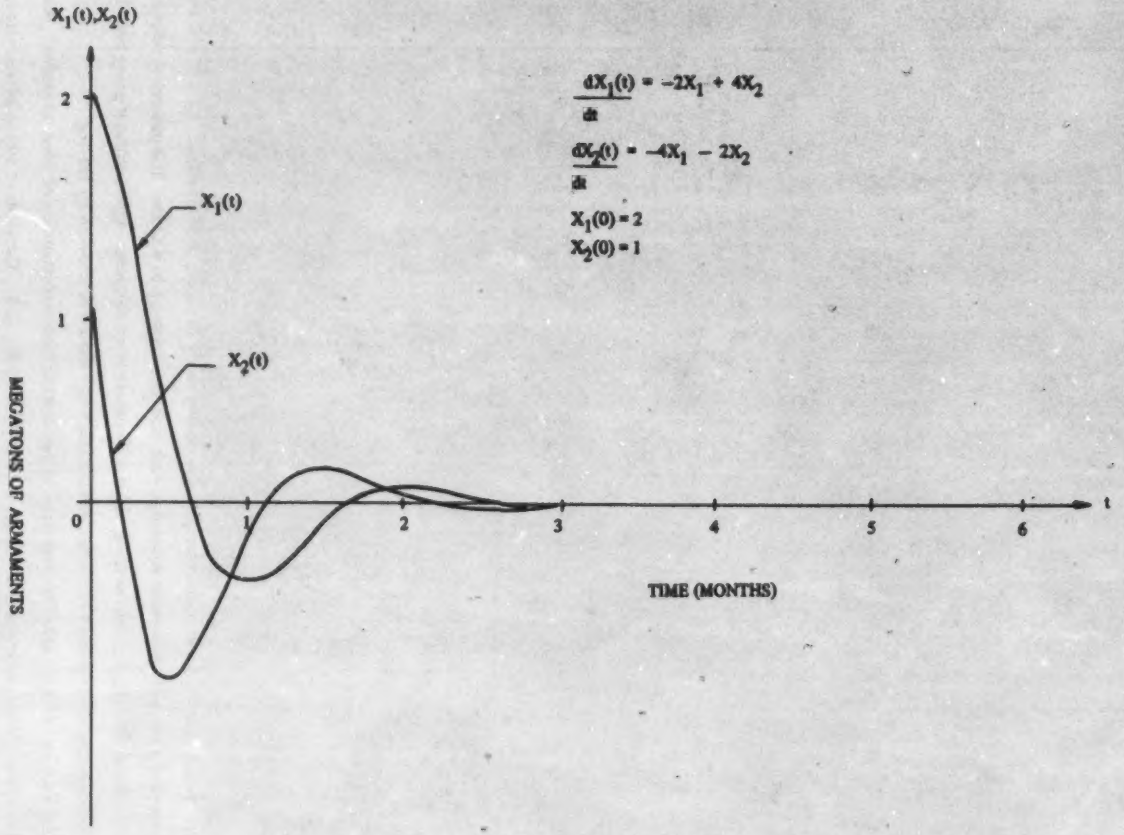


Fig. 27 Analog computer solution to oscillatory second order sample problem (the "arms race" problem). Note agreement between these results and the solid lines on Fig. 26.

CHAPTER 5

CLOSED LOOP FEEDBACK SYSTEMS

Introduction

In this chapter the concept of feedback is defined and investigated. Feedback is ubiquitous in technical, societal and biological interactions; it plays a role in systems control, economics, ecological balances and population growth to name just a few areas. Starting with the definition of feedback and closed loop systems and a brief historical survey, the chapter continues with some generalized analytical results to explain the rationalization and the basic behavior of feedback systems. Following this is the detailed analysis of one specific feedback system. The chapter ends by introducing some of the diverse fields in which closed loop feedback operations may be discerned.

Definitions

Imagine that a system, here described by G, is actuated by an input signal E which in turn produces an output signal C. Such a system can be represented in the block diagram of Fig. 28. At this point we do not have to be concerned with the internal operation of the system which we can treat as a "black box" except to note that the system operation will define a specific and unique relation between input and output. In the figure, the lines with arrows indicate the unidirectional flow of the E and C signals which might take the form of forces, voltages, currents, temperatures, dollars, men, material, etc. for different systems. The block includes the system descriptor* G which operates on (amplifies) the signal E to produce the signal C. G may be a numeric multiplier, a dynamic operator such as the characteristic equation of Eq. (50) (as seen in the next section), or some statistical operator to cope with random variations in the signal E. The equation describing Fig. 28 is given by

$$C = G \cdot E \,. \tag{80}$$

Fig. 28 An open loop system.

This system is called an open loop system since there is no connection from the output C back to the actuating signal E, i.e., there is no feedback. If this is to be a purposeful system where the actuating signal E is to produce a specific desired result, then we note that there is no assurance that this will actually occur. The system continues to operate on E to produce the output C regardless of whether this is at the desired level or not. As an example of an open loop system, consider a coin operated washing machine. Once the system is actuated, the machine goes through its cycles independent of whether clothes are being cleaned or not (or even present). Another example of an open loop system is the typical traffic light control system, which is preset to follow a prescribed timing pattern without accounting for the actual magnitude of the traffic flow in any direction.

---

* G is usually called the system transfer function, transferring the input signal E to the output C.

To overcome the disadvantages of open loop operation, we next investigate the possibilities of feeding back the output signal C to compare it with some desired level of the output, which is called the reference input R, and then actuating the system based on the difference between the fed back and the reference signals. In order to accomplish this, we will have to include a summing element which takes two signals (actuating and feedback) as inputs and produces an output which, in the case of negative feedback, is their difference. (Earlier, in our discussion of analog computers, we saw that a summing op-amp accomplishes this for voltage signals.) Fig. 29 presents the schematic representation of such a summing operation. This is described by the equation

$$E = (R - C). \tag{81}$$

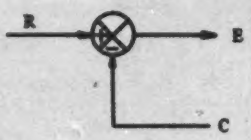


Fig. 29 A summer.

In practice, the output signal C must be measured before it is fed back, and frequently this measurement requires some form of operation which changes (or transduces) the output signal C to some other fed back quantity B, as shown in Fig. 30. In this case, the fed back signal B is equal to (H•C) and the equation that describes Fig. 30 is

$$E = (R - B) = (R - H \cdot C) \tag{82}$$

where H is known as the feedback transfer function.

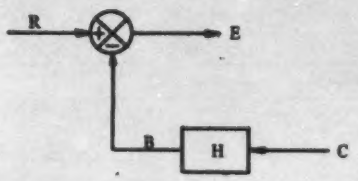


Fig. 30 Feedback measurement.

As an example of a feedback system, consider a highspeed electric passenger train which is to be run at a predetermined speed (corresponding to R). The actual system component to be controlled is the power to the traction motors to produce this speed and in practice this is achieved by having the engineer change the position of the power level. If the actual speed is below the desired speed, he advances the lever while he will move it back if the train runs too fast.

This system could also be automated by having the train speed measurement, C, translated into the motion of a wiper arm on a variable electric resistance which produces a certain voltage, B. This must then be compared to a voltage setting reflecting the desired speed, R. The actuating signal, E, would then be the difference, R – B, and this error signal would be utilized to speed up or slow down the traction motors by varying the power supplied to them through the necessary circuitry.

First of all, it should be noted that both B and R must be directly comparable and, in this case, both consist of voltage signals of essentially the same level. If we can make the assumption that the system behavior is linear and that the voltage produced by the wiper motion is proportional to the recorded speed, then the feedback transfer function, H, must have the dimensions of volts/unit of train speed or, perhaps, volts/miles per hour.

Secondly, the error signal may be quite small – of the order of a few volts – producing a signal whose power level is perhaps less than a fraction of a watt. On the other hand, the system to be controlled – the traction motors – may require power levels measured in kilo- or even mega-watts. The low power error signals are then used to control large amounts of power which are supplied by an external source. In the case of our example, this would be a third rail or the overhead wires.

An actual automated train system would, of course, have to be more complex since the desired reference speed would have to vary with the roadway. A reference speed of 100 miles per hour could be appropriate for a straight section of track but would have to be reduced ahead of a curve or before passing through a yard with many switches. In this case, an external signal would have to be fed to the train changing the reference speed. This must be done well enough in advance so that the train has time to respond to the new setting.

Consider, as another example, the thermostatically controlled home heating system. Small variations in room temperature are transduced into the motion of a bimetallic helical element. This motion, which is a low power signal, is next translated through a mercury switch into a system which can carry enough power to actuate the burner. Again, power amplification in the system is obviously needed. From an overall perspective, small thermal signals in the room are utilized to control large amounts of thermal energy flow in the fuel supply system.

Accordingly, the feedforward element characterized by G must inherently contain high power amplification. This idea is generally true in control systems. Small actuating effects (voltages, displacements, forces, etc.) usually control events involving large power levels.

We may also note from the previous discussion that the dimensions of H depend on the particular system selected. For instance, in a continuously varying temperature control system, the output signal measured in degrees F may have to be transduced into a voltage signal so that H would have the units of volts/degree. On the other hand, if H is to serve solely as a voltage multiplier (transducing a small voltage to a larger one), then its units would be volts/volt and it would be dimensionless and a pure numeric. In the simplest case, H would be both dimensionless and of magnitude one. The system is then referred to as a unity feedback system.

Combining Figures 28 and 30 produces Figure 31, the generalized feedback control system. It is important to note that the actuating signal E in Fig. 31 will be zero whenever the output exactly matches the reference input, i.e., when B = R. Further, it should be noted that a zero value for E does not necessarily imply a zero value for the output controlled value, C. A zero value of E merely states that the output will not change from its present operating condition. Furthermore, if the expected changes in operating conditions are relatively small, it is usually possible to linearize the system to operate about the nominal or desired operating point as was discussed in Chapter 2. We will assume from now on that this is possible, and all subsequent analyses deal only with small changes from a norm. Thus the powerful tools of linear analysis can be used.

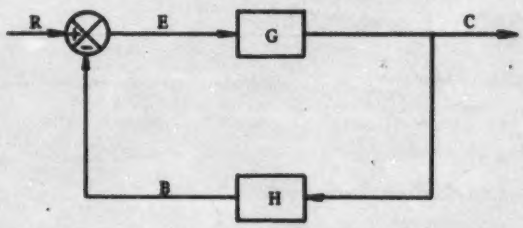


Fig. 31 A general feedback control system.

Substituting Eq. (82) into (80) yields

$$C = G\cdot(R - H\cdot C),$$

or, rearranging,

$$C\cdot(1 + G\cdot H) = G\cdot R \ .$$

Solving for the ratio of the output C to the reference input R leads to the output/input ratio of the system

$$\frac{C}{R} = \frac{G}{(1 + G\cdot H)} \ . \qquad (83)$$

which is called the closed loop transfer function. Expressing Eq. (83) in words, the closed loop transfer function is the ratio of the feedforward open loop transfer function G divided by (one plus the product of the feedforward transfer function G and the feedback transfer function H). This result is general and invariant as long as we are discussing negative feedback systems. For positive feedback systems the denominator is (1 – GH). Such systems have stability problems, as discussed in the next section, and are not generally useful for control purposes.

The fundamental importance of the feedback concept must be kept sight of through all of the detailed analysis which will follow. By utilizing feedback in a closed loop system which compares the actual controlled output with some desired level of the output (the reference input), a system is constructed which automatically accomplishes the desired objective. Thus, for example, in a traffic light control system, radar sensors could be mounted at busy intersections to sense the traffic flow and vary the timing pattern of the lights to avoid extensive delays in any one direction. In the case of a clothes washing machine, we might conceive of a "dirt sensor" (usually our eyes) which determines how long the machine should be left on. In either case, the addition of feedback creates a more desireable system.

Historical Background

The concept of feedback is an old one, going back at least 2,000 years to Greek and Roman water works and water clocks, in which the flow of water into the system was regulated by maintaining a constant level of water in a reservoir. In medieval times the concept of feedback was ingeniuously employed to keep windmills facing into the wind. The entire windmill was mounted on rollers so that it could rotate about a vertical axis. A vertical wind vane at the back of the windmill caused the whole top to turn when the vane was not lined up with the wind direction. By mounting the windmill blades on the opposite side of the wind vane, it was a simple and automatic matter to keep the blades turned into the wind and rotating at maximum speed.

James Watt gained renown in the late eighteenth century for improving the steam engine, and for applying the concept of a fly-ball speed governor to insure the safe operation of steam engines at regulated speeds. Since the position of the spring loaded fly-balls controlled by centrifugal forces which are directly proportional to the engine speed, their position is used to actuate a steam admission valve. Watt was thus able to prevent engines from over-speeding and exploding. If the engine tended to speed up, the fly-balls began to move out, which in turn closed the steam valve and reduced the speed. Conversely, if the engine tended to slow down, the fly-balls moved in to open the steam valve and speed up the engine. Watt's development proceeded by trial and error and without the benefit of theory or analysis.

A major development in feedback control occurred in the mid-1920's when the concepts and theory of feedback were applied to the ship-steering problem. Through the utilization of gyro-compasses and power actuators, ships were steered automatically by having an actuating signal proportional to the difference between the actual and desired ship heading angle amplified and fed to the power actuators controlling the rudder. About the same time, the instrumentation industry was starting to develop highly accurate null-balancing measuring devices which utilized the feedback concept. These compared a signal to be measured with a calibrated continuously variable standard signal. When the two balanced (nulled), the variable setting was read. Also, at about the same time, the process control industry started the development of numerous sophisticated feedback flow and pressure control devices used in various manufacturing processes. Such development continues today, particularly utilizing large-scale on-line digital computers for direct digital control (DDC).

In the mid-1930's the communications industry introduced the feedback concept into the development of feedback-stabilized amplifiers for long-distance voice communications. While doing so, many important theoretical developments emerged. The Nyquist stability theorem and the two Bode theorems, named after men at Bell Telephone Laboratories who developed them, were fundamental to subsequent advances. During World War II, the concepts of feedback control were applied to advanced automatically controlled weaponry. For example, radar and anti-aircraft guns were linked together to produce rapid and accurate gunnery systems. Much of this development was accomplished at the Lincoln Laboratories of MIT, some of it triggered by the theoretical work of Norbert Wiener. He also coined the word cybernetics (from the Greek word for helmsman) to describe the general feedback concepts used in information handling.

Following World War II, the theoretical developments were rapidly transferred to industrial applications, particularly to machine-tool control and automation. More recently, feedback control systems have also been applied to such diverse fields as space travel, bio-control, transportation planning, urban management, and many others.

Introduction to Feedback Control System Analysis

Consider the specific case of Fig. 28 and Eq. (80) when the open loop feedforward transfer function G is a pure amplifier gain of amplitude K (i.e., the output C is just K times the input E). If the gain of the amplifier were to change by 10%, then it should be clear that the output C will also change by 10%. To mathematically formulate this, let ΔC be the change in C and ΔK be the change in K. Then, since $\Delta C = (\Delta K)(E)$ and $C = (K)(E)$, their quotient yields

$$\frac{\Delta C}{C} = \frac{\Delta K}{K}, \tag{84}$$

or, the % change of the output is equal to the % change of the feedforward transfer function.

In contrast, now consider the closed loop case of Fig. 31 and Eq. (83). By a few algebraic steps*, it is shown that the closed loop equivalent of Eq. (84) is

$$\frac{\Delta C}{C} = \frac{\Delta K/K}{(1 + K \cdot H + \Delta K \cdot H)}. \tag{85}$$

Comparing Eqs. (84) and (85), it is seen that the percentage change in the controlled output ($\Delta C/C$) for the closed loop case is equal to the percentage change of the controlled output for the open loop case ($\Delta K/K$) divided by the quantity $(1 + K \cdot H + \Delta K \cdot H)$. If this latter quantity is large (say much larger than one), then the percentage change in the controlled output has been made largely insensitive to changes, ΔK, in the feedforward transfer function, K.

Taking a somewhat different approach, we can separately show that a feedback system can be made insensitive to feedforward parameter (e.g., gain) variations. Consider Eq. (83) by letting G = K and assume that $(1 + K \cdot H)$ is much larger than one. This requires that $K \cdot H$ itself is much larger than one, so that the equation can be approximated by

$$\frac{C}{R} \approx \frac{1}{H}. \tag{86}$$

In other words, when the gain K of the amplifier is very high, such that the product $K \cdot H$, is much greater than one, then the closed loop output/input ratio depends only upon the (inverse) characteristics of the feedback measuring device, and not upon the

---

* In the equation

$$C = \frac{G}{1 + GH} \cdot R$$

suppose that a deviation ΔC appears due to a change ΔG. Then,

$$C + \Delta C = \frac{G + \Delta G}{1 + GH + \Delta G \cdot H} \cdot R$$

and, by substituting the first of these equations into the second

$$\Delta C = \left[\frac{G + \Delta G}{1 + GH + \Delta G \cdot H} - \frac{G}{1 + GH} \right] \cdot R$$

$$= \left[\frac{(G + \Delta G)/G}{(1 + GH + \Delta G \cdot H)/(1 + GH)} - 1 \right] \cdot C$$

or, after algebraic manipulation,

$$\frac{\Delta C}{C} = \frac{(\Delta G/G)}{1 + GH + \Delta G \cdot H}.$$

Equation (85) follows by setting G equal to K and ΔG to ΔK. (This approach is taken from page 325 of Control, by Takahashi, Rabins and Auslander, published by Addison-Wesley, 1970.)

feedforward characteristics of the main power amplifying device. This result is general and is of fundamental importance in understanding closed loop feedback control concepts. The approximation of Eq. (86) suggests that in a highly accurate, proportional control system we should utilize feedforward power elements of high gain, and that we need feedback measuring devices of high accuracy. Therefore we should spend most of our money and energy on obtaining highly accurate small feedback measuring devices, whereas the large feedforward power elements may be inaccurate and inexpensive.

Unfortunately, there exists a very real practical limit to the amount that the feedforward gain K can be increased in order to validate the approximation of Eq. (86). As the gain is pushed higher, inherent time delays between the reference signal R and the output C of Fig. 31 start entering into the operation of the system. To understand the potentially dangerous consequences of such time delays, consider the case where the signal R is varying sinusoidally and the signal C accordingly varies sinusoidally at the same frequency. Due to the time delay, the output will be out of phase with the input. If the gain K is high enough to cause a time delay such that the output is at least one half cycle out of phase with the input, then the sign of the feedback becomes inverted. What was negative feedback now becomes positive and what was supposed to have been an error signal now shows up as an additional positive driving element. As a result of such positive feedback, the actuating signal begins to grow, since each time the signal traverses the loop it adds (due to the combination of negative feedback and phase shift due to time delay) to the previous signal level. As a result, the system output C becomes continually larger and the desired control function is destroyed. Such a situation is termed loop instability.

As an example of such an instability process resulting from high gain, imagine yourself driving an automobile with an excessively high steering ratio where small steering wheel changes result in large tire movements and directional changes. Suppose that a sudden disturbance, such as a car pulling out of a driveway just ahead of you, causes you to move the steering wheel suddenly by a large amount. Your objective is to avoid the car while staying on the road, i.e., to move slightly over and then to return the steering wheel rapidly into a near straight position. However, the various time delays inherent in the system (your reaction time, the inertia of the car, delays in the braking system) may easily cause a delay by half a cycle or more in the desired oscillatory signal (turn out and turn back straight). As a result, the car may begin to swerve wildly. With high gain (here a high steering ratio), very large feedback signals can be initiated which, instead of being corrective, may lead to increasing amplitude oscillations in the car's direction.

Associated with each system is a maximum level of feedforward gain which ensures stable system operation under normally encountered inputs. Feedback control system design thus evolves into a trade-off between a gain high enough to insure system accuracy and a gain low enough to insure relatively stable operation. The details of this trade-off design procedure are an important subject in texts on feedback control theory.

To further explore the differences in behavior between open and closed loop systems, we next investigate what happens when the feedforward transfer function G is no longer a pure gain, but also involves dynamic reactions between the actuating signal and the output. For instance, there may be a time delay between the actuating signal E and the output C due to the response characteristics of the system. We know from Chapter 3 that a time delay could be produced by a system representation which involves a first order differential equation and we can introduce this into the system dynamics by

$$T \cdot \frac{dC(t)}{dt} + C(t) = K \cdot E(t). \qquad (87)$$

In this equation, K is an amplifying gain which multiplies the actuating signal E(t), and T is a constant of the particular system under study which plays the role of the system time constant discussed in Chapter 3.

Rewriting Eq. (87) in operator notation

$$(T \cdot D + 1) \cdot C(t) = K \cdot E(t), \qquad (88)$$

or, after rearranging,

$$\frac{C(t)}{E(t)} = \frac{K}{(T \cdot D + 1)}. \qquad (89)$$

This equation is the open loop transfer function, G, for this system. As indicated in Fig. 28 and Eq. (80), G relates the actuating

signal, E, to the output, C. Note that the transfer function for a dynamic system (that is, one described by a differential equation such as Eq. (87)) involves the operator D (defined earlier as a short hand notation for the differentiation operation d( )/dt).

Dynamic system transfer functions, like Eq. (89), are exactly equivalent to system differential equations inasmuch as we arrived at this transfer function by starting with the system differential equation (Eq. (87) in this case). This process is reversible, so that given a system transfer function, we may arrive at the system differential equation. Thus, for example, assume that only Eq. (89) is known. We could then multiply both sides of Eq. (89) by the product of the denominators (i.e., cross-multiply). The result is Eq. (88). Finally, by replacing the short-hand D notation with its equivalent differention operation, the system differential equation (Eq. (87)) is derived. While this procedure has been demonstrated only for one specific example, it is entirely general and will work for any linear system transfer function.

Now assume that this same open loop input-output system is modified by adding unity feedback (H = 1), in order to create a closed loop feedback system. (A specific example of this operation will be considered in the next section.) The closed loop input-output ratio will be given by Eq. (83), where Eq. (89) may be subetituted for G in that equation. This leads to

$$\frac{C(t)}{R(t)} = \frac{G}{1+G} = \frac{\dfrac{K}{(T\cdot D+1)}}{1+\dfrac{K}{(T\cdot D+1)}} = \frac{K}{T\cdot D+1+K} = \frac{\dfrac{K}{1+K}}{\dfrac{T}{1+K}\cdot D+1} ,$$

or,

$$\frac{C(t)}{R(t)} = \frac{K'}{(T'\cdot D+1)} \tag{90}$$

where

$$K' = \frac{K}{1+K} \tag{91}$$

and

$$T' = \frac{T}{1+K} . \tag{92}$$

By comparing Eq. (90) to (89), it is seen that the nature of the closed loop system dynamics is similar to those of the open loop system, since each is a first order differential equation of the form of Eq. (87). However, after closing the loop, T and K in Eq. (87) are replaced by T′ and K′ of Eqs. (91) and (92). These seemingly small changes lead to significant changes in the system operation. For one thing, the amplifier gain, K, of the open loop system becomes K′ = (K/1 + K) when the loop is closed. Thus, for high open loop gain values of K (as required for small errors), the closed loop system will have an amplifier gain of essentially unity. If high gain between the reference input to a system and its output is a prerequisite of satisfactory performance (for either power or accuracy requirements), then in closed loop operation a pre-amplifier might be necessary to condition the reference input signal and bring it up to required power levels. A well known application of this situation occurs in hi-fi units, where the feedback power amplifiers have low gain.

However, a far more important change has occurred in closing the loop; what was an open loop time constant of T time-units has become, in closed loop operation, T′ = (T/1 + K). Thus, as the open loop gain K is increased, the effective closed loop time constant gets smaller, and the system responds to input changes more rapidly. These results are general – one important effect of closing the loop around a system is to speed up its time response, but this occurs at the expense of lowered system gain.

To study the actuating signal E in more detail, let us reconsider Fig. 31. This signal may be written as

$$E = R - B = R - H\cdot C = R - H\cdot G\cdot E$$

or,

$$E(1+G\cdot H) = R$$

or,

$$\frac{E}{R} = \frac{1}{1+G\cdot H} . \tag{93}$$

This result, called the error response or error transfer function, is particularly informative when we let H = 1. For this unity feedback case, E = R – C is a direct measure of the difference between the controlled output C and the desired level R of the controlled output.

In other words, the signal E directly measures the system error. An ideal control system is one that keeps C in correspondence with R at all times and for all signals R, so that the system error E is always zero. One way to influence the system in this direction is to design it so that the magnitude of G·H is always much greater than one. Then E/R in Eq. (93) will always be a small number. For example, with G = K, a pure gain and with unity feedback, Eq. (93) indicates that the error E will tend towards zero as the gain K gets very large. Physically this also makes sense, since the larger the gain K in the feedforward box of Fig. 31, the smaller need be the error E before some actuation occurs to eliminate that error.

It would appear from the discussion of the previous paragraph that a logical design conclusion would be to construct closed loop systems with as high a feedforward gain as possible. Unfortunately, as stated previously, the dynamic stability requirements of the system precludes this choice since increasing the gain also increases the tendency towards unstable system operation. In practice, a reasonable balance must therefore be found between the accuracy and the relative stability requirements of the system.

An Example of a Typical Feedback Control System

Imagine, as shown in Fig. 32, a water tank of constant cross-section A (square meters) fed by a faucet at the top and with a leak at the bottom. The variable flow of E(t) (cubic meters of water per second) out of the faucet is controlled by the faucet handle setting. The variable leakage flow of L(t) (cubic meters per second) out of the tank bottom depends on the water level C(t), the variable to be controlled in this system. The higher the level of C(t), the larger will be the leakage flow L(t). While the actual relationship between C(t) and L(t) is non-linear – it is a quadratic relationship as found by application of Bernoulli's principle – we will assume a locally linearized version of the relationship as

$$\mathcal{L}(t) = \frac{c(t)}{K} \qquad (94)$$

where c(t) is a change in C(t) about an equilibrium level, C_{eq}, and $\mathcal{L}(t)$ is a change in L(t) about the equilibrium leakage flow, L_{eq}. K is an empirically determined constant with dimensions of seconds per square meters to insure dimensional consistency in Eq. (94). In the actual non-linear system, measurements would show that large variations in $\mathcal{L}(t)$ would not be proportional to c(t). Furthermore, small local variations for different values of C_{eq} might require different values of K to represent the system. This is just to serve as a reminder that linearization of an actual non-linear system limits the analysis to both small variations and to a specified equilibrium value.

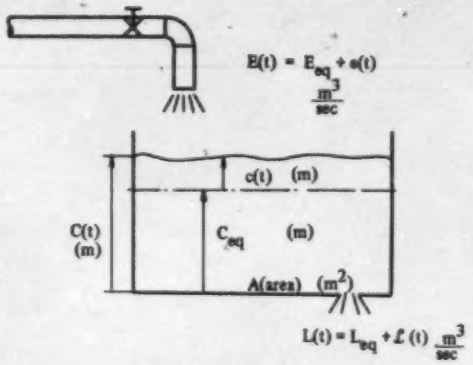


Fig. 32 A leaking tank

It is perhaps surprising to learn that, within limits, an equilibrium level does indeed exist in Fig. 32. As the faucet is opened wider to increase E(t), the change in level c(t) will increase until a new equilibrium level C_{eq} is reached which forces an increased leakage flow L(t) to just balance the increased input flow E(t). For a given tank with a given leakage hole at the bottom, each different faucet setting (E_{eq} value) leads to a different equilibrium level. A simple experiment in a bathroom sink confirms this observation.

As the faucet setting is changed by an amount e(t), the level C(t) dynamically changes from one value to another. We next

study the dynamics of this change c(t). The equations will all be written in terms of the change variables ($\mathcal{L}$(t), e(t) and c(t)) about the equilibrium condition. To make this study, we note that water can neither be created nor destroyed and must be continually accounted for. By applying this concept of conservation of matter, a continuity expression is derived which states that the net flow of water into (or out of) the tank shows up as an accumulation (or depletion) of water in the tank with passing time.

The net flow of water into or out of the tank is given by

$$\text{Net Flow} = e(t) - \mathcal{L}(t), \tag{95}$$

where a positive value of this net flow indicates an accumulation of water in the tank. This is given by

$$\text{Water Accumulation} = (A)\cdot\frac{dc(t)}{dt}, \tag{96}$$

where A is the tank cross-sectional area (in square meters) and dc(t)/dt is the time rate of change of the tank level. This last term is seen to be the "velocity" of the liquid level, with a positive value indicating a rising level. Multiplying this velocity by the cross-sectional area gives precisely the cubic volume swept out by the moving level which must be equal to the net flow into the tank. Accordingly Eqs. (95) and (96) must be equal. This is merely another way of stating continuity and results in

$$e(t) - \mathcal{L}(t) = (A)\cdot\frac{dc(t)}{dt}. \tag{97}$$

Substituting Eq. (94) into Eq. (97) to eliminate the leakage flow $\mathcal{L}$(t) (which is not of primary interest at present), and rearranging,

$$(A)\cdot\frac{dc(t)}{dt} + \frac{c(t)}{K} = e(t). \tag{98}$$

Next multiplying through by the constant K, letting the product K·A (which has units of time) be replaced by a new constant T, and utilizing operator notation,

$$(T\cdot D + 1)\cdot c(t) = K\cdot e(t). \tag{99}$$

This is a first order differential equation with forcing input e(t). From Chapter 3, we know that a step change in e(t) of magnitude E in^3/sec will cause an exponential change in c(t), where the nature of the exponential change is described by the system time constant of T time units. c_{ss}, the steady-state value of the change in c(t), is found from Eq. (99) by letting D go to zero (i.e., once at equilibrium the level is no longer changing and the time rate of change of the level is zero), so that $c_{ss} = K\cdot E$. This gives the change in the level, so that the absolute value of the new steady-state level is the old equilibrium level plus this change.

After some simple algebraic manipulations, equation (99) may be re-written as

$$\frac{c(t)}{e(t)} = \frac{K}{(T\cdot D + 1)}. \tag{100}$$

Comparing this result to Eq. (89), it is seen that Eq. (100) plays exactly the role of the open loop transfer function G between the actuating signal of the changed faucet flow e(t) and the system water level change c(t). Thus, Fig. 28 and Eq. (80) pertain directly to this open loop water tank problem.

As a small digression, consider the effect of closing the leak, so that L = 0, or, equivalently, K = ∞. By dividing the numerator and denominator of the right hand side of Eq. (100) by K, and replacing T by A·K, Eq. (100) becomes

$$\frac{c(t)}{e(t)} = \frac{1}{A\cdot D}, \tag{101}$$

or,

$$c(t) = \frac{1}{A}\int_0^t e(t)dt. \tag{102}$$

Eq. (102) indicates a pure integration of e(t) to produce the change in level c(t). Thus, a step increase in the faucet flow rate of magnitude E would cause a level that would increase linearly with time as given by (E(t)/A). This checks our physical understanding

of what would happen to the water level in the tank if the leak were plugged and the flow rate were kept constant.

Let us next consider how to make the open loop water tank a closed loop control system. One obvious possibility is to keep our hand on the faucet and our eye on the water level in order to maintain the water level at some desired height irrespective of any disturbances acting on the system. If, for example, a bar of soap accidentally fell into the tank (slightly raising the water level), we would manually turn off the faucet a small amount to bring the level back down to its desired value. But it is very tiring to keep a constant look out for possible disturbances, so we next seek a means of automatically controlling the liquid level in the face of any system disturbances.

As one possibility, consider the float and valve arrangement of Fig. 33. The float F is constrained to move vertically by the guideways J. As the float moves up due to a disturbance in the water level (i.e., as the liquid level changes by an amount c(t) above the equilibrium level C_{eq}), the bar APB pivots about its fulcrum P and causes the valve V to move down, in turn causing the total inlet flow rate E(t) to decrease, providing negative feedback. This decrease will cause the level to start dropping back towards the equilibrium setting. If the disturbance c(t) were negative, i.e., a drop in the level below C_{eq}, then the valve would be opened, causing an increase in E(t), which would in turn cause the level to start rising back towards the equilibrium setting.

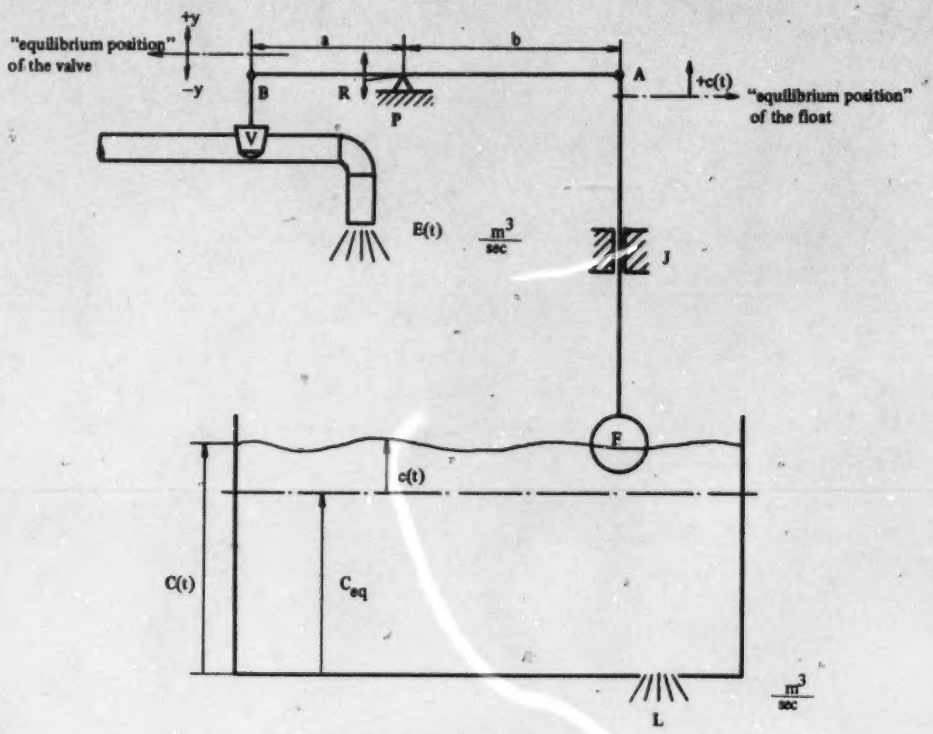


Fig. 33 A leaking tank with automatic level control.

The equilibrium setting C_{eq} can be changed by moving the pivot point P up or down. This corresponds to changes in the reference input R(t)* in Fig. 31 and Eq. (83). To quantify the above discussion, observe that (for R = 0)

$$y(t) = -\left(\frac{a}{b}\right) \cdot C(t) , \tag{103}$$

where a and b are the distances from the ends of the bar to the pivot point as shown in Fig. 33 and the minus sign is the manifestation of the negative feedback described above. Further, we assume a linear relationship between the valve motion y and the change in

* To be consistent, R(t) should be defined as $R(t) = R_{eq} + r(t)$, but this nicety will be omitted.

input flow, e(t), about that level which just maintains C_{eq}, or

$$e(t) = M \cdot y(t) , \tag{104}$$

where M has units of (square meters/sec). When y(t) is zero, so is the change in flow, e(t), about the equilibrium input flow rate. For a negative y(t), a drop below the input equilibrium flow rate results. At this point, the defined signs should be carefully checked against the physical behavior of the system as depicted in Fig. 33 to insure the understanding of negative feedback (i.e., an increase in c(t) causes a decrease in y(t) and vice-versa). If the valve had been attached to the bar between points A and P, then an increase in c(t) would have yielded an increase in y(t) (i.e., positive feedback), resulting in an unstable system. In this case, any change in the level c(t) would result in a dynamic response that would grow without bound . . . or at least until the tank overflowed or drained itself empty.

Eq. (103) gave the valve motion as a function of float motion c(t) when the reference setting R(t) was kept at zero. Next consider the case when c(t) is zero, but R(t) is not (i.e., a level set-point change), causing the bar to pivot about point A. Simple trigonometry yields

$$y(t) = + \frac{(a + b)}{b} \cdot R(t) . \tag{105}$$

The total valve motion for any combination of motions c(t) and R(t) is given by the sum of Eqs. (103) and (105), or

$$y(t) = + \frac{(a + b)}{b} \cdot R(t) - \frac{a}{b} c(t) . \tag{106}$$

Substituting Eq. (106) into Eq. (104), the change in input flow rate e(t) due to both reference input changes R(t) and controlled level changes c(t) is found to be

$$e(t) = + M \cdot \frac{(a + b)}{b} \cdot R(t) - M \cdot \frac{a}{b} \cdot c(t) . \tag{107}$$

Eq. (107) completely specifies the behavior of changes in the input flow rate due to changes in either the reference input or the controlled level. On the other hand, Eq. (100) completely specifies the changes in the controlled level c(t) as a result of changes in the input flow. The two equations are drawn onto one block diagram as shown in Fig. 34. It is seen that the feedback transfer function H is equal to the quantity M(a/b) (square meters/sec). The fed back signal, B(t) = H·C(t), is thus checked to have the correct units of cubic meters per second. B(t) is dimensionally consistent with the signal R′(t), which is [M·(a + b)/b times R(t)], where R(t) is measured in meters. Since the coefficient K in the feedforward box has the dimensions of sec/square meters as discussed before, it is seen that the entire block diagram of Fig. 34 is dimensionally consistent.

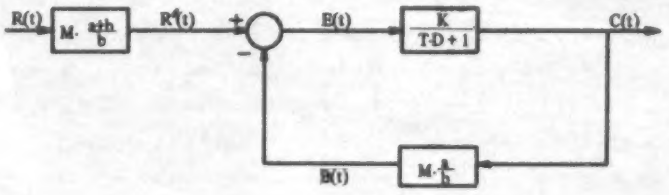


Fig. 34 Block diagram for a leaking tank with feedback level control.

The discussion of a unity feedback control system in the previous section is, for this system, simply a special case of Eq. (107) and Fig. 34. If the valve gain M and the level ratio (a/b) happen to be such that their product is unity, then Fig. 34 becomes a unity feedback system and the ration c(t)/R′(t) is given exactly by Eq. (90). By making K very large (i.e., cutting down on the leakage), it is seen from Eq. (92) that the closed loop system time constant can be made very small. The quantity M·(a + b)/b which multiplies R(t) to produce the signal R′(t) plays the role of a system "pre-amplifier" as discussed in the previous section.

Finally, consider the case when the leakage is zero, so that Eq. (101) applies. Letting the quantity M(a/b) be represented by

some number N_1 and $M\cdot(a+b)/b$ by some other number N_2, then the closed loop feedback water level control system is represented by the block diagram of Fig. 35. Applying Eq. (83) to Fig. 35, where the R in Eq. (83) is the $R'(t) = N_2R(t)$ of Fig. 35,

$$\frac{C(t)}{R'(t)} = \frac{C(t)}{N_2\cdot R(t)} = \frac{\dfrac{1}{A\cdot D}}{1+\dfrac{N_1}{A\cdot D}} = \frac{1}{A\cdot D + N_1} = \frac{\dfrac{1}{N_1}}{\left(\dfrac{A}{N_1}\right)D + 1},$$

or,

$$\frac{C(t)}{R(t)} = \frac{(N_2/N_1)}{\left(\dfrac{A}{N_1}\right)D + 1}. \qquad (108)$$

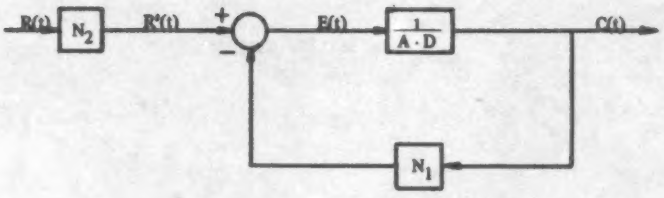


Fig. 35 Block diagram for the tank with zero leakage.

Eq. (108) and the block diagram of Fig. 35 are applicable to the system of Fig. 33 for the case when the leakage flow, L, is zero. This is representative of the filling of a toilet tank. To speed up the system response (decrease the filling time), the system time constant (A/N_1) in Eq. (108) would have to be decreased by appropriate redesign of the tank, pivoted bar, or the filling valve. However, the situation of Fig. 33 is applicable to any liquid level control system where there are inlet and outlet flows. Such situations frequently occur in the process controls industry.

The techniques of analysis of this simple first order system are quite general and allow for complex, higher order systems which are handled in similar fashion. The complexity enters in longer equations, but not necessarily in additional concepts.

Most important, this simple but real example serves as an analog for numerous other important situations If we think of the level of the water tank as being analogous to population level in a country, then inlet and outflow rates play the role of combined birth and immigration rates and combined death and emigration rates respectively. The feedback float mechanism is replaced by census data and the inlet valve is replaced by natural pressures on births and immigration. If these conditions hold, the simple water tank analysis is then analogous to a simplified population dynamics problem. At the same time, it should be realized that this population dynamics model may be over-idealized since not all the controlling variables nor their interactions and time delays have been accounted for. Obviously, only a more complete analytical model can lead to a full description of the actual complex population study. However, it should be pointed out in passing that even very complicated dynamic systems can often be represented by nothing more than non-linear first order mathematical models.

A similar examination of the governing factors in simplified studies of lake eutrophication, traffic accumulation, predator-prey ecological balances, body hormonal responses and of a host of other problems will show that a straightforward first order feedback system will lead to considerable insight and, frequently, to a satisfactory analysis.

Closure

The concepts of dynamic feedback control systems that have been developed to this point consist of far more than a methodology for analyzing a class of problems. Rather they comprise a perspective on technology, nature, and society which enables us to discern feedback loops in operation and, by applying some of these concepts, to better understand some of the problems we

continually face.

Thinking back to the armaments race problem of Chapter 4, that problem is immediately recognized now as two coupled first order feedback systems. The feedback is evident from the analog computer patching diagram of Fig. 21-(c). The feedback mechanism takes the form of either "spying" on the competing nation's armament level, or being politically sensitive to one's own.

The whale survival problem, discussed in Chapter 3, is recognized as a first order feedback system where the feedback occurs naturally through the effective sensing of current population levels to create new birth and death rates. Also in that chapter, the first order traffic level problem is seen to comprise a natural feedback loop in that the level of vehicles in Manhattan effected the rate of change of that level.

The human body is a mass of complex and inter-related self-regulating feedback loops. To isolate just one, consider the body temperature regulation process. When we are overheated, we perspire and pump our cooling blood at a faster rate. When we are cold, we shiver and internally metabolise stored carbohydrates. This system works within remarkably small tolerances about the normal body temperature of 98.6° F, and it is hard to imagine an industrial controller working as long and as well. Other self-regulating feedback loops within the human body maintain chemical balances, blood pressure, vertical equilibrium, etc. The sum total of all such processes is known as homeostasis, or, roughly, "keeping things as is."

Feedback mechanisms also operate in the economic and social milieu in which we exist. Witness the feedback coupling between demand, supply and price which can lead to some highly unpredictable dynamics in the economic sphere. There have been recent attempts at applying dynamic feedback system concepts to the complex problems of our urban society.* While it is premature to judge the complete applicability of the results, these studies have focused attention on the various factors which contribute to dynamic interactions in society and have shown that attempts to alleviate problems based on "intuition" are often counterproductive since our intuition is not capable of allowing for complex dynamic interactions with variable time delays.

All of these studies have one facet in common: real systems are described by coupled first order differential equations involving both positive and negative feedback loops. To this extent, the material presented in these brief introductory notes can therefore serve as a starting point in the analysis of the real and complex dynamic problems of our everyday life.

---

* J. W. Forrester, Urban Dynamics (1968) and World Dynamics (1972), M. I. T. Press, Cambridge, Mass.

Appendix A

POLYTECHNIC INSTITUTE OF BROOKLYN

ORSA DEPARTMENT

SA 200. "CONCEPTS OF MODERN TECHNOLOGY"

SYLLABUS – SPRING, 1972

WEEK	LECTURE TOPIC	NOTES
1	Organization & Introduction, Analog and Digital Computation	
2	Problem formulation & causality. Criteria, alternatives, constraints, trade-offs, and objectives.	
3	Introduction to dynamic systems and analog simulation.	
4	Dynamic systems modeling.	Book choices due.
5	Engineering, ecological and environmental dynamic system examples.	
6	The concepts of feedback.	Quiz no. 1 (In Lab.)
7	Technological and non-technological feedback examples.	Project choices due.
8	The Forrester approach.	
9	Urban, social and economic examples	Interim Project Report Due.
10	Modeling and Simulation — Continuous Systems Modeling Program with Bioengineering Example (Guest Lecture).	
11	Operations Research Methods — Introduction to Linear Programming, Queueing Theory and Game Theory (Guest Lecture).	Quiz no. 2 (In Lab.)
12	Decision Making — Economic Decision Models, Quality Control and Simulation (Guest Lecture).	Book reports Due.
13	Technology Assessment (Guest Lecture).	Final Project Reports Due.
14	Seminar on Course Readings.	

Freshman Workshop Assignments

Digital and Analog Computer Simulations

Digital Computer Simulations

The programs listed below (developed by Prof. L. Braun of Stony Brook) are available on tape for student use via a PDP–8 terminal. The student is required to run fourteen of these programs, which are written in BASIC. Each program requires student "data" input in an interactive mode so that the actual result will depend on the choices made by the student. A log book of program listings, results and a summary page for each program must be handed in by the student.

BIOLOGY

SIMULATION	TOPIC
BUFLO	An exercise in resource management, based on the life cycle and population patterns of the buffalo.
GENE1	A simple genetics experiment to demonstrate the statistical nature of the Mendelian Laws.
HARDY	The Hardy–Weinberg Principle. (Population Genetics)
LOCKEY	Competitive inhibition; the Lock and Key Model of enzyme action.
MALAR	Control of a malaria epidemic.
PH	How pH affects enzyme activity.
POLUT	Effects of pollutants on different bodies of water and the effectiveness of certain anti-pollution measures.
POP1–POP5	A series of simple population models.
PYR1	A sampling exercise, on a simulated farm pond, for investigating population size.
RATS	Rodent control.
STERL2	Control of an insect population by pesticides and the use of sterile males.
TEMP	How temperature affects enzyme activity.

PHYSICS

SIMULATION	TOPIC
CHARGE	Millikan's oil drop experiment. (Discreteness of the electric charge.)
FRANCK	Franck–Hertz experiment. (Discrete energy states in electrons.)

(continued on next page)

PHYSICS (continued)

SIMULATION	TOPIC
GRAV	How the distance between masses and the amount of mass affect the gravitational attraction. (Based on Cavendish experiment.)
OPTIC	Optical Bench
PRJCTL	Projectile motion.
QUANTA	Photoelectric effect.
REACTR	Operation of a nuclear reactor.
SCATR1–4	Use of alpha particle scattering to determine the atomic structure.
SLITS	Young's double slit experiment. (Interference of light waves.)

SOCIAL STUDIES

SIMULATION	TOPIC
BALPAY	U. S. balance of payments.
DECIDE	Stream pollution with economic and social factors included.
ECON1	A simplified model of the U. S. economy.
HIRISE	Growth and development of an urban area.
MARKET	Simulation game engaging two companies in a one-product competition.
POLSYS	How local government responds to different attempts of the community to influence its decisions.
QDM	Quantifying decision-making.
STOCK	A chance to play the stock market.
URBCOM	Conflict between different factions of an urban community.

Analog Computer Simulations

These are primarily selected from "The Man-Made World" (1971 hard cover edition) and include:

Buffalo Harvesting (pp 186 – 192)
Models of Exponential Growth (pp 192 – 193)
Vehicle Response (pp 237 – 238)
Resource Allocation of Man Power (pp 238 – 239)
Falling Body Simulation (pp 241 – 242)
Dynamic Population Model (p 243)
Periodic Signals (pp 289 – 291)
Water Pollution Model (pp 291 – 295)
Feedback Control (pp 333 – 334)
Water Level Control (pp 335 – 336)
Spring Deflection Simulation (pp 338 – 340)
Home Heating Model (p 341)
Aerodynamic Stability (p 381)
Boat Docking (pp 418 – 419)
LEM Landing Simulation (pp 420 – 421)

Appendix C

POLYTECHNIC INSTITUTE OF BROOKLYN

ORSA DEPARTMENT

SA 200. "CONCEPTS OF MODERN TECHNOLOGY"

READING LIST -- SPRING 1972

Required Text:

The Man-Made World, McGraw-Hill, New York, 1971.

Report List:

Each student is responsible for reading two of the following books and writing a one-page précis and summary report on each one. These reports are due on April 25th. Student book choices are to be cleared with the instructor to minimize duplication. Copies of these reports will be distributed to the class one week before the May 9th seminar discussion on this reading list. Books preceded by an asterisk are available from the instructor on a loan basis.

Alsop, Joseph and Alsop, Stewart, We Accuse, Simon and Schuster, New York, 1954. (An accounting of the purported facts behind the Oppenheimer case.)

*Altman, Stanley M. (Editor), "System Science and Cybernetics --Special Issue on Urban and Public Systems", I.E.E.E. Transactions, Volume SSC-6, Number 4, October, 1970. (A collection of 10 timely papers.)

*Black, Guy, The Application of Systems Analysis to Government Operations, Frederick A. Praeger, New York, 1968. (Principles of System Analysis.)

*Blumstein, Alfred, et al (Editors), System Analysis for Social Problems, Washington Operations Research Council, Washington, D.C., 1970. (Papers on city management, human resources, transportation, and environmental qualities.)

*Braden, William, The Age of Aquarius, Quadrangle Books, Chicago, 1970. (Technology and the cultural revelation.)

*Brzezinski, Zbegniew, Between Two Ages, The Viking Press, New York, 1970. (America's role in the technitronic era.)

*Calder, Nigel, Technopolis, Simon and Schuster, New York, 1970. (Social control of the uses of science.)

*Chacko, George K. (Editor), The Recognition of System in Health Service, Operations Research Society of America, Arlington, Va., 1969. (Collection of papers on organization of health services, manpower, hospital care, and the role of technology.)

*Churchman, C. West, The Systems Approach, Dell Publishing Co., New York, 1969. (A critical appraisal of system analysis compared to other approaches.)

*Forrester, Jay W., Urban Dynamics, The M. I. T. Press, Cambridge, Mass., 1969. (A system approach to the diagnosis of current urban problems.)

* , World Dynamics, Wright-Allen Press, Cambridge, Mass., 1971. (Interrelates population, industrialization, natural resources, food, crowding, and pollution, to present an exploratory theory of man in the world ecology cast in a computer simulation model.)

Kipphardt, Heinar, In the Matter of J. Robert Oppenheimer, Hill and Wang, N.Y., 1968. (A play -- somewhat fictionalized.)

*Melman, Seymour, Our Depleted Society, A Delta Book, 1965. (An in-depth study of the utilization of our national resources.)

Mumford, Lewis, The Pentagon of Power, Harcourt, Brace, Jovanovich, N.Y., 1970. (The Myth of the Machine --part of a continuing series on science and society.)

*Odom, Howard T., Environment, Power and Society, Wiley –Interscience, New York 1971. (An analytical study.)

Oppenheimer, J. Robert, The Open Mind, Simon and Schuster, New York, 1955. (Eight collected lectures on atomic weapons, the scientist and society.)

Oppenheimer, J. Robert, Science and the Common Understanding, Simon and Schuster, N.Y., 1953. (6 BBC lectures on science from a lofty perspective.)

*Ramo, Simon, Cure of Chaos, David Mc Kay Co., New York, 1969. (System Analysis as an alternative to uncontrolled growth.)

*Scott, Ellis L. and Robert W. Bolz (Editors), Automation and Society, center for the study of automation and society, Athens, Georgia, 1969. (Twelve papers given at a recent symposium.)

*Snow, C. P., The Two Cultures and a Second Look, Cambridge University Press, London, 1963. (A famous lecture and some follow-up retrospective comments on science and society.)

*Snow, C.P., The New Men, Penguin Books, London, 1959. (A novel about atomic physicists and their conscience.)

*Snow, C.P. Science and Government, Harvard University Press, Cambridge, Mass., 1961. (A series of 10 lectures on the role of the scientist in forming governmental policy - in the case bombing strategy during W. W. II)

*Toffler, Alvin, Future Shock, Random House, New York, 1970. (A current bestseller on the social problems involved with rapid technological change.)

*Travias, Irene (Editor), The Computer Impact, Prentice-Hall, Englewood Cliffs, N.J., 1970. (A series of papers on effect of computers on government, the economy, privacy, law, education, and society.)

*U.S. House of Representatives, Science and Technology and the Cities, U. S. Government Printing Office, Washington, D. C., (A compilation of papers by noted experts.)

*Vonnegut, Kurt, Jr., Player Piano, Avon Pocket Book, 1952. (A novel on the engineer's role in the futuristic society.)

* , Science and Technology and the Cities, U. S. Gov't Printing Office, Washington, D. C. (Complete testimony before the Committee, February 4, 5, 6, 1969.)

* , Technical Information for Congress, U. S. Gov't Printing Office, Washington, D. C. (Expert testimony on how Congress can keep abreast of, and evaluate technical developments.)